FINAL REPORT
CIFA CONTRACT – FA 4814-04-0011

Voice Stress Analyzer Instrumentation Evaluation

Harry Hollien, Ph.D.
James D. Harnsberger, Ph.D.
Investigators
IASCP, University of Florida
Gainesville, Florida

February 28, 2006

TABLE OF CONTENTS

I. INTRODUCTION

I-A. Background

I-A-1. Perspective

It is well known that the speech signal contains features which can be used to provide information about a human speaker. "Voice identification" is based on one of these sets of features as numerous speaker specific phonatory properties have been discovered (see among many others: Hollien, 1990, 2002; Hollien and Schwartz, 2002; Kuenzel, 1994; Nolan, 1983; Stevens, 1971). Another such area involves the detection of alcohol intoxication as it is reflected in voice and speech. Here too, a substantial amount of research has been reported which provides intelligence about these relationships (see among many others: Chin and Pisoni, 1997; Hollien et al, 1998, 2001 a and b; Klingholtz et al, 1988; Pisoni and Martin, 1997). Human emotion (including psychological stress) constitutes yet a third domain where behaviors can be detected in voice (see among others: Cummings and Clements, 1980; Hicks and Hollien, 1981; Hollien, 1980, 1990; Scherer, 1981, 1986; Williams and Stevens, 1972).

The neurological bases for the relationships described above also are reasonably well established. That is, since the speech act represents the output of a number of high level and integrated neurological systems (sensory, cognitive, motor), it appears appropriate to assume that the process may reflect a variety of other conditions. Specifically, since the oral production of any language involves the use of multiple sensory modalities, high level cognitive functioning, complex cortical processing and a large series of motor acts (see among others: Abbs and Gracco, 1984; Netsell, 1983), it is logical to predict that even more subtle operations -- such as the detection of deception and/or truth from speech and voice -- also would be possible.

I-A-2. Detecting Truth-Deception-Stress from Voice

There is no question but that a device which could detect the presence of truth, deception and/or stress from voice/speech analysis would be of great value to intelligence groups. Several systems that are purported to do so currently exist. As a group they are referred to as voice stress analyzers (VSA); they will be identified by this term (i.e., VSA), even though their function may or may not be based on a "voice stress" protocol.

To date, research on several of these systems has ranged from mixed to somewhat negative. Some authors suggest that these devices might possibly detect stress -- at least in certain circumstances (see Brenner and Branscomb, 1979; Brockway et al., 1976; McGlone, 1975; Van der Carr, et al., 1980). However, most research has not supported this position (e.g., Horvath, 1979; Cestaro and Dobbins, 1994; Cestaro, 1996; Janniro and Cestaro, 1996; Meyerhoff et al., 2000; see below also). On the other hand, it can also be said that, to date, none of these instruments have been afforded a full evaluation. Most investigators simply have not controlled their procedures at a level which would permit the necessary information to be generated (for example, see Brenner et al., 1979; Heisse, 1976; Lynch and Henry, 1979; O'Hair et al., 1985). Some have not assessed a sufficient number of variables or have carried out research of too limited a scope (see Greaner, 1976; McGlone et al., 1974; Leith et al., 1983). The focus of others has been, perhaps, too narrow (Inbar and Eden, 1976; Leith et al., 1983; Shipp and Izdebski, 1981) or they have carried out only limited laboratory studies (Brenner and Branscomb, 1979;

McGlone, 1975). Still others have reported projects that were somewhat restricted, even if reasonably well controlled (Haddad et al., 2002; Hollien et al., 1987; Horvath 1978, 1982). Finally, some researchers have limited their effort only to field studies (Barland, 1975; Kubis, 1973) and, no matter how competently they were carried out, this approach does not provide the fundamental information necessary. Only Nachshon and Feldman (1980) attempted both laboratory and field studies. However, even here, the effort lacked breadth and sufficient control.

When addressing the challenge of properly evaluating VSA devices, it also is necessary to specify the types of experiments that would be carried out. So far, the most common approach has been to conduct research by means of a class of experiments which can be described as "simulated field" studies. The reason for doing so appears to be the desire of the investigator to determine if the system will work under "real life" conditions. Moreover, there are relevant individuals (see, for example, Lykken, 1981) who argue that laboratory experiments are simply "games" and since they are "unrealistic," they provide little-to-no useful information. The counter arguments to that position are that 1) field research ignores the need for basic system assessment under *controlled* conditions, 2) it does not include events necessary for the proper determination of system operation, 3) it does not exclude debilitating external variables and 4) knowledge is lacking on the speaker's actual behavioral states. Thus, little can be gained from the "field experiment only" approach because sufficient information does not become available about the basic product provided by the system. Even more important, since reasonable controls cannot be applied to research in that milieu, it is not possible to determine if the information obtained from field research is valid. This difference of opinion creates a very real dilemma. In response, the questions may be asked: Is it possible to conduct "laboratory" experiments that are realistic enough to provide useful data about the VSA systems; is it possible to conduct field experiments that are precise enough to generate valid data, and should both be included in a project such as this? Our response here is in the affirmative. Further, we have developed a three level model as a guide (see section I-D). However, the tasks completed for this contract (section I-B) and the more general goals of this research program (section I-C) are listed prior to specification of the model.

I-B. Task

The project task was to evaluate two specific deception detection systems in a highly relevant and highly controlled manner. The two systems were the National Institute for Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA) and Nemesysco's Layer Voice Analysis (LVA), distributed in the United States by V. The devices were tested in a double-blind study rather than one that included both an operator and an on-scene event involving human subjects. It was through the use of this paradigm that the systems themselves, apart from the operators' abilities or use of non-system information, could be evaluated in a thorough and impartial manner. The project was completed as tasked and in its entirety; the results are provided in Section III below. In the course of completing all tasks, additional relevant research was conducted – as were procedures which included the use of highly experienced operators provided by the manufacturers. These several additional subprojects are reported along with the project

results to enable a larger understanding of the sensitivity of these two devices to deception in speech.

I-C. Goals

This project had two primary goals: 1) The development of effective experimental procedures and speech samples suitable for the evaluation of the capacity of voice stress analyzers (VSA) to detect truth, stress and deception from voice and speech and 2) Apply these procedures in the evaluation of two VSA devices, NITV's CVSA and Nemesysco's LVA. Until this project was conducted, neither system had been adequately assessed primarily because (see above) 1) past studies had not been extensive enough to do so; 2) their experimental procedures had not been adequately structured and controlled; 3) stress and deception had not been measured independently to determine their actual relationships and 4) carefully designed laboratory and field experiments had not been conducted – especially within a single study. The research team at the University of Florida's Institute for Advanced Study of the Communication Processes (IASCP) addressed the limitations of prior work by structuring a model, then conducting research with the general objectives of:

1. Detecting stress in speech/voice
2. Detecting deception in speech/voice
3. Detecting truth in speech/voice
4. Detecting stress combined with deception in speech
5. Conducting both laboratory and field research
6. Using rigorously controlled procedures
7. Studying a reasonably large number of appropriate samples produced by two large populations of speakers
8. Applying the procedures developed for objectives 1–7 in thorough and impartial tests of two VSA devices.

The research approaches outlined above highlighted an important problem in the evaluation of VSA systems: such devices typically rely on the effects of stress on the acoustic properties of speech as a direct indicator of deception. However, stress and deception can be separated and, furthermore, numerous behavioral states can give rise to stress-based changes in voice. Therefore, stress and deception had to be examined separately, then in combination, within the same study to model their relationship. The speech materials collected for these purposes were drawn from two types of experiments, *laboratory-based* and *field-based*. The laboratory studies provided for basic system assessment under controlled conditions. The resulting dataset also included real-time measures of speakers' actual stress levels (as based on both physiological and psychological measures) which provided verification that a speech sample was produced under stress. The field-based data provided an evaluation of the voice stress systems with presumably greater external validity.

I-D. Model

The model generated in support of this and related projects is three-tiered in nature. The first level involves highly controlled laboratory experiments and evaluations. The second level is focused on both 1) simulated field and 2) real field research, but

where only low levels of control and verification are possible. The third level involves actual field experiments -- often referred to as "real life" studies -- where data (and the results of system evaluations, of course) are obtained under conditions of modest-to-high level control and validation. All three approaches lead to the development of test vehicles designed to evaluate equipment (in this case, VSA devices). Yet more importantly, they lead to the generation of information about those basic parameters which signal stress, deception and truth in voice and speech. A brief review of these three approaches or "levels" follows.

I-D-1. Laboratory-based Studies

Research of this type requires high levels of subject and procedural control and that all behavioral and experimental conditions are verifiable. In this case, all experiments must be double-blind, the stimuli shown to induce the desired behavior, subject's responses validated and so on.

Briefly, our basic or core study provided a range of scenarios that varied truth, stress and deception with jeopardy and speaker intention; they included:

1. Truthful, unstressed utterances (baseline material of several types)
2. Deceptive speech produced under low jeopardy
3. Deceptive speech produced under high jeopardy
4. High-stress truthful utterances
5. Simulated high-stress speech (but where the subject actually was experiencing low stress)

I-D-2. Field Studies

As stated, the field research involves either simulated field procedures or actual cases involving interrogation but where neither the subjects' (often suspected criminals) stress level nor guilt are verifiable.

Initially, we planned to draw speech materials from several sources; they were to be of several types.

1. Neutral, unstressed utterances (baseline material)
2. Low jeopardy lies (no rewards/punishment)
3. Modest to high jeopardy lies

These field materials were to be obtained under relatively realistic conditions. That is, recordings were to be obtained from actual interrogations conducted by police and military officers. Our caveat would be that only those recordings where the presence of falsehood could be reliably determined would be used.

One field study of the first type plus a small investigation of specialists in signal detection were carried out in this area. (Note: These studies were not tasked but were completed by the investigators).

I-D-3. Actual Field Research

Studies here would be of the type where our teams would be present when crimes *potentially* could be committed. One example would be where inmates of a prison were interrogated (and recorded) as to whether or not they had recently taken certain illegal

drugs. Whether they had or had not done so would be verified by blood/urine tests. In turn, the recording could be employed in both basic and system evaluation research.

No research at this third level was contracted for or carried out under the auspices of this contract.

I-D-4. Application of the Obtained Materials

Ultimately, the speech materials collected in the laboratory and field-based studies (Levels 1 and 2) were used by two *evaluation groups* in assessing the accuracy of two VSA devices:

1. Two IASCP team members who had received formal training by two VSA manufacturers, NITV and Nemesysco.
2. Two-to-three trained representatives provided by each VSA manufacturer. Note: this section constituted an upgrade of the specified task.

A third group, phoneticians who specialize in signal analysis, evaluated one VSA device, CVSA. It was not appropriate to have the phonetician operator group evaluate LVA because the optimal method devised by the IASCP team to evaluate LVA did not require an operator (see III-C-2).

II. METHOD

II-A. The Laboratory Level Research

As stated, the primary objectives of this project were to carry out highly controlled research that would at once be 1) impartial to all sides of the prior VSA controversies – i.e., those which led to the need for this research and 2) rigorous enough to address questions concerning the validity and sensitivity of the systems involved; in this regard the equipment alone was to be evaluated first. Hence, it was clear that a large and diverse sample of subjects (i.e., speakers) was required; one that encompassed men and women who varied in both age and socioeconomic background. It was critical that the recorded speech samples involve high jeopardy and that the stress level of the speakers during production be independently determined. In addition, truth, deception and stress had to be examined as independent variables primarily because the detection of stress itself is important since it may provide important information for intelligence purposes. Moreover, it may be as easy as or easier to detect from speech than is deception.

II-A-1. Protocols

Details of the basic or core study follow.

II-A-1-a. Subjects and Recording Procedure

78 Adult volunteers, both male and female, were first screened for suitability re: inclusion in the study. Their ages ranged from 18 to 55 years and they represented a diverse demographic sample. All potential subjects were screened by the co-investigator psychiatrist who excluded those individuals with relevant medical conditions that could

be exacerbated by stress or who had a past history of psychological trauma; many other potential exclusionary mental and physical health criteria also were assessed.

Subjects were recorded reading materials under various conditions of stress while producing truthful statements and lies. All recordings were made in a quiet (but "live") room with laboratory quality microphones coupled to 1) a DAT recorder, 2) a digitizer attached to a desktop computer and 3) and an analog cassette recorder. Digital audio-video recordings of each subject were made during all experimental runs. The video cameras were fixed and focused on the subject's upper body.

II-A-1-b. Stress Level Controls

Five procedures appropriate for the measurement of psychological stress were administered either simultaneously with the audio and video recordings or once during or after each experimental procedure; they were: 1) Two tests of anxiety/stress level (based on self-reports) administered after each experimental condition, 2) a saliva test also taken at that time, and 3) body response evaluations of galvanic skin response (GSR) and pulse rate (PR) collected during each procedure and monitored throughout the entire subject run. The anxiety/stress tests (based on self-reports) consisted of an "emotion felt" anxiety checklist (see Appendix A.1) and a modified version of the Hamilton test (Maier et al, 1988; see Appendix A.2). The cortisol level (saliva) tests were accomplished by Salimetrics LLC, No. 5100 Cortisol Tests. GSR and pulse were measured continuously using the BIOPAC Systems, Model MP-150.

II-A-1-c. Speech Samples

Seven different types of speech samples were obtained from each subject-speaker. They were elicited by six procedures and during baseline calibration following a familiarization process:

- Baseline calibration: The subject read a standardized phonetically-based (unstressed) truthful passage, namely the Rainbow Passage
- Procedure 1: The subject read a neutral (unstressed) passage which was truthful.
- Procedure 2: A passage was used wherein the speaker produced a lie while not experiencing significant stress.
- Procedure 3: The subject uttered untruths under jeopardy (see below).
- Procedure 4: Truthful speech was uttered at a relatively high stress level (i.e., stress induced by mild electric shock).
- Procedure 5: Untruths were uttered both under high jeopardy (as in Procedure 3) along with fear induced by the administration of electric shock (see below). It was by this procedure wherein jeopardy was created by two stimuli applied simultaneously.
- Procedure 6: Truthful utterances were produced but where the subject simulated speaking under stress while not actually stressed.

II-A-1-d. Speech Sample Characteristics

The speech samples recorded for these six procedures were carefully designed. First, they were extensive enough to provide a reasonable repertoire for all types of VSA evaluations. Specifically, they were relatively long and varied enough to permit operators the opportunity to make valid decisions regardless of the device being tested. To this end,

each passage consisted of 5-7 sentences. A 17-25 word neutral phrase or sentence was embedded within each of them in order that no language cues about the condition being experienced were inherent within target-utterance syntax. An example of such a phrase is: "This is a position I am very comfortable with because I have thought about it for a while and it makes sense." Note that it is not specific to any particular topic. The use of neutral content phrases prevented system operators from being exposed to language-based clues as to the nature of the speaking condition.

Procedure 1: After reading the baseline calibration passage a number of times, the subject read a truthful passage; that is, he or she was permitted to become familiar with, and then read, a passage about an unemotional topic.

Procedure 2: The low-stress deceptive utterances were created in a similar fashion except false statements were spoken. Examples: "I now live at 3120 Northwest 38th Drive," and so on, or "I find most jazz compositions to be rather nonmusical," (plus related text).

Procedure 3: These materials consisted of untruths produced under high jeopardy. All subjects had been selected from groups that were known to hold very strong personal views about some issue (such as religion, politics). They were recorded uttering statements that contradicted these strong views, all the while under the impression that their friends and/or other peer groups would hear their performance. In addition, subjects were instructed to produce these lies in a speaking style that strongly suggested that they believed them. These instructions both supported the jeopardy speaking condition and permitted the capture of a reasonably useful repertoire of appropriate speech.

Procedure 4: The stress-only procedure consisted of subjects reading truthful material, namely statements with which the speaker agreed but was not particularly passionate about. For this procedure, he or she was conditioned to respond to the highest level of mild electric shock that could be tolerated. They were told that they would receive a shock whenever they produced the neutral sentence appearing in the middle of the passage. As stated, the equipment employed in the conditioning process was a BIOPAC Systems, Model MP-150 with an electro-stimulus conditioning unit (STM100C). After conditioning, electric shock was administered during the initial run of the procedure and in any subsequent runs wherein the subject failed to show significant signs of stress during the production of the experimental passage (as determined by the physiological measures of GSR and pulse rate). The electric shock was administered using two tin electrodes set 30mm apart within a watertight acrylic bar. The electrodes were attached to the inside forearm of the subject's non-dominant arm.

Procedure 5: This experimental condition combined elements from procedures 3 and 4. Specifically, the Procedure 5 materials consisted of harsh lies produced under the high jeopardy of being heard by confederates, just as did Procedure 3. In addition, the threat of receiving electric shock was also present, just as it served as a stressor in Procedure 4. Therefore, Procedure 5 contained two stressors and was used to elicit lies under the highest degree of psychological stress possible under laboratory conditions.

Procedure 6: The samples here were obtained after the subject was coached to produce a truthful passage in a manner reflecting how they might speak under conditions of significant stress. The subject was allowed to repeat this procedure until he/she and the experimenter agreed that utterances had been produced that were different from their normal speech (and presumably "reflected stress").

The procedure number specified above reflects the original structuring of these materials. However, after pilot research was conducted, a different order of presentation was developed and used in testing most subjects. That is, the final order within a trial grouped procedures that involved stress together (e.g., Procedures 3, 5 and 4) followed by those that did not involve stress (Baseline plus Procedures 1, 2 and 6). The detailed description of the testing sequence follows.

1. After reporting to the lab and giving informed consent, participants (i.e., potential subjects) completed the "Subject Information Form."
2. The project's psychiatrist and medical director (Camillo A. Martin, M.D) screened subjects using a series of questions concerning those aspects of their background that might make them inappropriate for the study. General screening questions covered the following topics: 1) history of psychiatric disorders, 2) history of heart conditions 3) other physical disorders, 4) current medication regimen, and so on. None of the subject's responses to these questions were recorded. They simply were used to include or exclude them from the study -- and to add an element of uncertainty to the session.
3. The subjects who qualified were:
    a. Seated in the testing room and had a head-mounted microphone (Shure SM-10A) fitted to them.
    b. The GSR and pulse rate sensors were then placed on two fingers of the non-dominant hand (later the electro-shock stimulator was placed on the subject's other arm, but only for Procedures 4 and 5).
4. Procedure 3 trials. Two or more runs were carried out with the subject producing different passages that were judged to be both offensive to his/her strongly-held beliefs and were entirely untruthful. The saliva test for cortisol and the two self-report tests were administered at the end of this procedure.
5. Procedure 5 trials. Calibration of the electric shock stimulus was carried out first. Up to three runs were conducted with different passages that were judged to be entirely untruthful and objectionable. Following these runs, the saliva test for cortisol and the two self-report tests were administered.
6. Procedure 4 trials. Up to three runs were made with different passages that were judged to be entirely truthful by the subject. The purpose of this procedure was to induce speech produced under the stress caused by the fear of electric shock. Again, the saliva test for cortisol and the two self-report tests were administered at the end of the procedure.
7. After the completion of the stressful procedures, subjects were debriefed as to the actual purpose and use of the materials elicited in Procedures 3 and 5. The transducer for administering shock was removed, and the subject was engaged in

conversation with the research personnel to set him/her at ease for the subsequent low stress procedures.

8. After a break, the subjects read a calibration or baseline passage (i.e., the Rainbow Passage). At the end of this calibration passage, the saliva test for cortisol and the two self-report tests were once again administered.
9. Procedure 1. This procedure involved producing an unstressed (neutral) truthful passage. The saliva test for cortisol and the two self-report tests were administered at its end.
10. Procedure 2. This procedure involved producing an unstressed deceptive passage on a topic that was not of direct interest to the subject – hence, a low stress lie. The saliva test for cortisol and the two self-report tests were then administered.
11. Procedure 6. This procedure was typically run two to four times to elicit a sample of simulated stress produced under low actual stress conditions (based on the physiological correlates being measured). It was also one where the subject imitated stress in voice, in the judgment of the PI or Co-PI. At the end of this procedure, the saliva test for cortisol and the two self-report tests were administered for the last time.

The use of the protocols described here enabled us to develop a practical database of speech samples; one that contained all of the linguistic information needed to test a variety of voice stress analysis products -- plus provide material for basic research. The speech materials also were verified as containing lies produced under jeopardy. The actual degree of psychological stress was quantified by the use of multiple converging measures. The final product of this protocol -- the speech materials that constitute the basic or Voice Stress Analysis (VSA) database -- represent a unique resource in the evaluation of current and future commercial voice stress analysis products.

II-A-2. Database development

Of the 78 human subjects who completed the protocol described above, 55 met the minimum criteria for inclusion in the VSA database. These criteria focused on the *shift in stress* as measured by both the physiological correlates that were continuously measured as well as the self-report scales collected after each procedure. All of these potential measures of stress (plus cortisol) were examined independently to determine whether or not they each showed a significant shift from the unstressed conditions (e.g., procedures for eliciting low-stress truthful statements, low-stress lies, simulated stress) to the stressed conditions (e.g., procedures for eliciting high-stress truthful statements, high-stress lies). Four of these metrics showed significant difference in the required direction (i.e., stressed samples > unstressed samples). They were galvanic skin response, pulse rate, the emotion checklist, and the modified Hamilton scale. One measure, cortisol, failed to show a significant shift in the required direction and was excluded from the composite measure of *stress shift*. Like other studies which have shown mixed results for cortisol levels, our measures here (averaged over two tests of all of the samples) failed to show a significant, or even systematic, difference in the anticipated direction between the unstressed and stressed conditions.

A review of the literature on its reliability as a physiological correlate to psychological stress confirmed that the results on cortisol testing vary considerably

across studies. That is, many researchers have found widespread individual differences within their testing (Bohnen et al., 1991; Bossert et al., 1988; Smyth et al., 1998), while others have demonstrated that significant changes in cortisol levels can occur consistently across large groups (Bassett et al., 1987; Nejtek, V.A., 2002). Aside from individual differences, researchers have also observed a connection between gender and cortisol levels. Males seem to exhibit a higher level of cortisol (compared to baseline) in anticipation of and during a specific stimulus, while females exhibit either a lower or unchanged level of cortisol in response to the same stressor (Frankenhaeuser et al., 1976; Kirschbaum et al., 1992). Such differing results with cortisol likely reflect large methodological differences between studies. For this study, it appears likely that cortisol level does not shift quickly enough to provide useful information for our rapidly changing experimental procedures, making it an inappropriate to serve as a physiological correlate to stress in our protocols.

Of the five potential stress correlates examined, four were ultimately included in the stress shift composite score: galvanic skin response, pulse rate, the emotion checklist, and the modified Hamilton scale. The greatest combined stress shifts (in both the physiological correlates as well as the self-report scales) were used to select a subset of the speech samples collected that ultimately constitute the VSA database. Specifically, overall stress shifts were computed by averaging all four measures after they had been converted to a common scale. Equal weighting was assigned to each in determining the overall shift. Given this metric, we were able to include a total of 48 subjects in the VSA database (out of the 55 who met the minimum criteria) whose stress level while lying was typically **more than double** their baseline stress level. (Baselines were calculated for individual speakers by selecting the procedure showing the lowest GSR and pulse rates during the entire procedure). Specifically, the mean overall stress shift observed across the 48 speakers selected was 141% (129% for male speakers, 152% for female speakers), with median, minimum and maximum shifts of 128%, 45%, and 392%, respectively. The resulting database, then, consists of 48 speakers, 24 male and 24 female, all of whom produced deceptive statements while under a significant degree of stress. A list of the individual stress shifts calculated for all subjects can be found in Appendix B.

Please note that the speech materials appearing in the VSA database consist of the middle (neutral) sentence from one passage re: each condition. It should be emphasized once again that this neutral sentence is embedded within the total sample. Hence, it has been shown to powerfully reflect the stress level being experienced by the subject even though it does not linguistically reveal the content of the passage.

The speech materials cited were organized into ten sets of thirty samples (five male and five female sets) with a total of 56 speakers employed across all ten sets (48 recorded under the protocol and eight recorded as foils). Four sets of the male subjects and four sets of females contain different speakers. A fifth set for each group was developed for reliability evaluations and draws subjects from the other four data sets. Thus, there are thirty samples within each of ten sets, 28 from tested subjects and two non-stress foils recorded by different individuals, for a total of 300 samples. The foils were recorded by male and female volunteers who had not participated in the experiment. The new foil-speakers read passages that were considered to be truthful for them but were used as lies for other subjects during the testing phase of the experiment. In summary,

each of the ten sets includes ten different subjects who produce from 1-5 truthful and deceptive utterances of the types described above.

II-A-3. Testing CVSA

The CVSA system was designed for testing single syllables, which can either constitute whole words or only parts of words. Most commonly, it is used to analyze live voice recordings of single syllable, single word responses, namely "yes" and "no." While it can be utilized for running speech, operator judgments of CVSA's processing of the speech signal are normally made on single syllables extracted from the longer speech samples. Therefore, we were required to extract single syllables from the neutral material found in the VSA database. For high stress deceptive and truthful samples, single syllables were drawn from Procedures 3, 4 and/or 5 – ones that occurred at the maximum in both physiological measures (GSR and pulse rate). As stated, these physiological maxima were obtained by first converting both sets of measures to a normalized, common scale and then combining them into a single dataset. The syllable selected on the basis of the stress maximum also had to meet three other criteria: 1) it could not exhibit an abrupt onset or offset of vocalization, 2) the articulation had to occur at a typical intensity level (no trailing voice effect at end of reading) and 3) it had to be produced with phonatory output in the modal (normal) register – i.e., no breathy samples were acceptable nor were those in the falsetto or vocal fry registers. Syllables at the physiological maximum were not selected if they did not meet all three criteria. Instead, the syllable nearest to that maximum and which met all three criteria was chosen. For the low stress deceptive and truthful samples, single syllables at the physiological minimum were drawn from Procedures 1, 2 or from the calibration passage (the Rainbow Passage), using the same methods and criteria as those used in selecting the high stress samples. Finally, the complete set of VSA syllables was randomized to ensure that no stress or deception information based on ordering was available to any CVSA operator.

It may seem that inordinate care was taken in preparing the samples for use by the CVSA operators. However, this approach was employed to ensure that, if CVSA was sensitive to any degree to deception or stress in speech, it would be measurable with the database we developed. Moreover, we also are aware of a second potential issue in testing CVSA with any speech samples. Specifically, it is possible that certain phonetic relationships might possibly complicate interpretation of the CVSA charts (see section III-A-2 for an explanation of CVSA operation). Related speech waveforms can show different patterns -- depending upon the particular vowels and consonants being spoken; variation in intensity further complicates these relationships. For example, the combination of /m/ and /a/, spoken with their typical intensity and duration could result in the misclassification of a truthful or unstressed speech sample as a lie or as stressed. Accordingly, we chose a variety of phoneme combinations to input into CVSA.

II-A-4. Testing LVA

The VSA database was also reorganized for use with the LVA system. First, LVA is not limited to analyzing single syllables, thus the full length samples drawn from our database could be employed. Second, the LVA device requires sentence-length speech materials at a minimum and also requires that a "balanced" portion of an individual's normal speech be added for calibration purposes. That is, LVA must first extract speech

norms for a given speaker in order to accurately classify a particular speech sample as deceptive or stressed (or as exhibiting some other cognitive or emotional state). Thus, to prepare the 300 speech samples for submission to LVA, all of them had to be individually paired with a section of the "Rainbow Passage" produced by the corresponding subject. The 300 pairs were then inputted as single digital audio (wave) files. The Rainbow passage (one from each speaker-subject) served as calibration material for that subject's experimental speech sample. Finally, the complete set of digital audio files for LVA were assigned random filenames (using an alphanumeric code) to ensure that no stress or deception information about the sample was available to any LVA operator.

II-B. Field Research

A single federal intelligence agency provided the project with a set of audio-video recordings of military trainees answering questions while undergoing SERE training (Survival Escape Resistance Evasion). The SERE program is a rigorous survival training program where the students are trained not to reveal any information when captured and interrogated by hostile forces.

II-B-1. SERE Study Characteristics

The particular SERE trainees that were recorded took part in a guilty knowledge study in which subjects were instructed to lie about several aspects of this training. The goal of the study was to detect lies embedded in a large number of truthful responses. In turn, subjects faced punishment if their lies were detected. Thus, they were lying under a substantial degree of jeopardy, although they did not face a severe immediate threat.

While being recorded on video-camera, the SERE subjects wore a Vivometrics "Life Shirt" that continuously recorded common physiological correlates of stress; included are metrics such as heart rate, breathing and blood pressure. In this case, the SERE subjects exhibited heart rates typically varying between 140 and 170 BPM, with 95 being the lowest value recorded. In contrast, their base heart rates were relatively low when they were at rest; they ranged between 48 and 52 BPM. Thus, it appeared reasonable to infer that the threat of punishment associated with this procedure resulted in a substantial elevation of stress levels during the interrogation.

II-B-2. Procedures

The materials received consisted of audio-video recordings of 26 SERE subjects on whom a research team had collected data. Of the 26 subjects available, only seven actually had produced deceptive statements. However, from this pool of speech materials, a SERE database was developed that includes a total of 56 utterances consisting of either a "yes" or a "no" response to a question. Given the limited duration of these responses, this database could not be used to evaluate LVA. However, they were highly suitable for testing CVSA.

The 56 utterances were organized into related sets of eight speech samples, six sets for the males and two for the females. Each set contains samples drawn from five subjects. Three of the utterances were produced by the primary SERE subject, two more of the utterances by other SERE subjects, and the final two samples were produced by "foils." The foil talker recordings were obtained from individuals working at IASCP; they were not involved with SERE in any way. Each of the eight sets contains three lies

and four truths. One lie and one truth were drawn from target stimuli, two truths and one lie were selected from non-target stimuli, and the final two samples are truthful foils. A target stimulus is defined in terms of a predetermined question the examiner has asked that particular subject. Other samples were randomly drawn from all of the question categories presented during the experiment (i.e., instructor, animal, aircraft type, letters, numbers, and security classification). Each set also varies by type of utterance; this means that some sets contain four 'yes' and three 'no' answers while other sets contain the opposite pattern. To code the utterances for use with CVSA, the letters assigned to each set were combined with an arbitrary number 1-56 (an example of a coded sample is 'A-42'). This coding system allowed for randomization in each group as well as ensuring that pertinent information about each sample would not be inadvertently disclosed.

III. RESULTS

III-A. Organization of the Results

This part of the report will be organized into two major sections -- plus conclusions. The first will focus on the several evaluations carried out on NITV's CVSA system; the second on Nemesysco's LVA device. In turn, each of these major sections will be divided into two related segments. The first, identified as Summary Results, will focus on the main findings/consequences of the research. The second -- which will be referred to as Technical Results -- will be longer. It will include comprehensive presentations and discussions of the findings plus the statistical analyses. It is in this section that the obtained data will be presented in a variety of ways.

III-A-1. The CVSA Analyses. -- General.

It must be remembered (see above) that the 300 samples for the CVSA analyses were drawn from the core or VSA data base in the form of single syllables. These samples were inputted into the appropriate laptop computer using its sound card and as directed by the manufacturer. Once all samples were inputted, two trained CVSA operators from the IASCP team classified each sample as "deceptive" or "non-deceptive." This judgment was based on the presence or absence of "blocking" in the CVSA charts. A third individual (the PI) maintained the key to the randomization of the samples and collected the operators' results. This procedure was to ensure that the integrity of the double-blind procedure was not compromised. Subsequently, three highly experienced operators, provided by NITV, traveled to the University of Florida and also classified these same samples as "deceptive" and "nondeceptive."

The second phase of this evaluation process occurred when the IASCP and NITV teams analyzed the SERE-based field materials. The procedures followed in this second set of experiments exactly paralleled those carries out for the first study.

Finally, four phoneticians, experienced in visual analysis of acoustic signals and related configurations "read" the two sets of charts (Core: N=300; SERE: N=56). They were provided only a short explanation of "blocking" (drawn from the CVSA manual) and a few samples of classic blocking and nonblocking. They were only asked to complete a set of forced choice "blocking-nonblocking" judgments. This group was recruited to ensure that the sensitivity of the device was understood with respect to the

operator's experience with the device. The NITV team represented the one with the greatest experience with CVSA. The IASCP team had been trained by the manufacturer and had received certification, although they possessed less experience with the device than the NITV team. Finally, the Phonetician team possessed no experience and received only minimal training in the form of instructions.

III-A-2. The CVSA charts

The CVSA charts can be described as two dimensional displays in which the duration of the speech signal is displayed on the horizontal axis; the information on the vertical axis is not defined. A sample pair of charts appears in Figure 1.

As stated, the left chart is supposed to display a voice recording in which psychological stress is present. Its gross shape would be referred to as "blocking" (in the CVSA training manual) due to its general rectangular form. The right chart displays a voice recording in which psychological stress is presumed to be absent. That is, the operator would judge that "blocking" is absent in this chart and do so on the basis of its more triangular configuration. Specifically, it appears to have a "peak," with the signal strength rapidly decreasing at onset and offset. Such charts would be classified as nondeceptive and unstressed. NITV states that blocking -- the single cue for stress that may be a product of deception -- results from the suppression of a natural "microtremor" in the muscles that control both the vocal folds and all other muscles employed in speech articulation. It is claimed that when this microtremor is suppressed, its acoustic byproduct -- referred to as the "inaudible frequency modulation (FM) component" -- is lost. In turn, this results in the appearance of "blocking" in the signal seen on a CVSA chart. When the subject is no longer under stress, the microtremor returns and blocking dissipates.

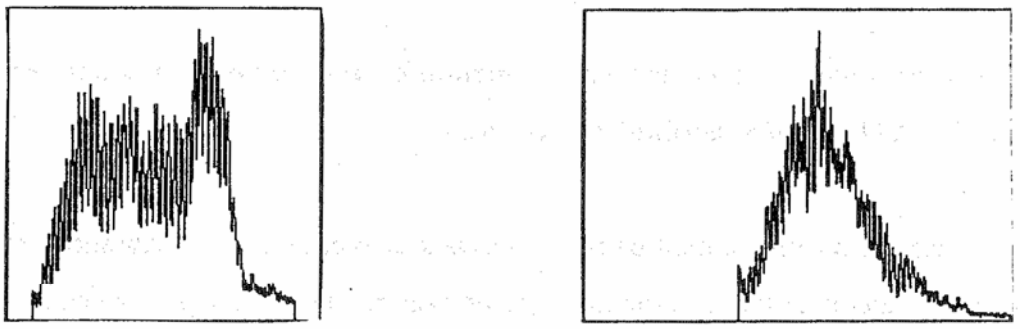

Figure 1: The left chart ("Less FM") shows "blocking" due to stress and/or deception. The right chart ("More FM") shows an absence of blocking, a pattern which would be interpreted as unstressed and/or not deceptive.

III-A-3 The LVA Research - General

Several relationships should be stressed at the offset. First, the LVA system is dealt with in a different manner than is the CVSA equipment. Longer passages are required; so is a calibration passage, uttered by that same subject. Once this pair is inputted, the relevant laptop (i.e., the one containing the proper LVA program) the only thing the operator must do is select that analysis procedure (from a rather extensive corpus of readouts) that he or she believes will best identify the emotion/condition being

expressed/intended by the subject. Accordingly, no interpretation of waveforms or waveform processing is necessary. The operator simply selects the output considered relevant and reports the results. Thus, the LVA analysis is conducted automatically without any operator "bias" or effects.

Second, while the IASCP team established a set of protocols for generating the data, the team from V (Nemesysco) was permitted to use any of set of the available sets they wish to apply re: their analysis (see Appendix C for the dialog carried out between the investigators and the LVA officials). In any event, the IASCP team followed a fixed set of protocols for the analysis of both stress and deception-truth. For stress, LVA's "JQ" parameter was used as it appeared most appropriate for assessing this behavioral state. Specifically, a sample was coded "stressed" if the mean JQ level across all relevant segments of a subject's speech sample was 35 or greater. For deception, the test sample was so coded if "Final Analysis" stated that "Deception was indicated in the relevant questions." (Please see section III-C for more details concerning the IASCP team's analysis).

Accordingly, the 300 double passages were inputted by one of the operators (trained to do so by LVA). The IASCP team worked together analyzing the data obtained on the basis of the above-cited protocols. The LVA team (two senior individuals -- both of whom are instructors at the training school), however, was permitted to discuss which of the readouts to use for each of the 300 samples. They were observed to use several different protocols in a number of instances. Hence, it can be said that the IASCP used a fixed set of protocols when making their decisions whereas the V team was not consistent in this regard; they could, and did, use a variety of readouts in their decision-making.

III-B. Summary Results

The results were examined by means of a number of techniques designed to explore the possibility that CVSA and/or LVA may be sensitive to stress and/or deception. In all approaches, four rates were calculated: *true positive*, *false positive*, *false negative* and *true negative*. The true positive rate (or "hit rate" in Signal Detection Theory), refers to the proportion or percentage of the time that deception or stress is said to be present when in fact it is actually present. True positive rates measure how often the device accurately classifies a deceptive utterance as deceptive. Equally important was the calculation of the false positive rates (also known as *false alarm rate* in Signal Detection Theory). They correspond to the percentage of times the signal is said to be present when in fact it is absent. They answer the question: "How often does the device classify a truthful utterance inaccurately as a deceptive one?" False positive rates **must** be compared with true positive rates in order to determine a device's sensitivity to deception or stress. An examination of true positive rates alone do not determine the accuracy of a device since a high true positive rate could be the product of either the device's actual accuracy or its bias to simply classify speech samples as deceptive regardless of the actual presence or absence of deception. A sensitive device would show true positive rates that are both high and significantly different from the false positive rates. A device that performs at chance with respect to deception or stress would show relatively equal true and false positive rates.

Finally, the false negative and true negatives rates were also determined (also known as the *miss rate* and *correct rejection rate*, respectively, in Signal Detection

Theory). False negatives occur when the signal is present but the detector (in this case, the device operator and/or the device's output) classifies it as absent. They represent inaccurate performance by the device and/or operator. True negatives are cases in which the signal is in fact absent and it is accurately judged to be absent (e.g., truthful speech samples that are classified by a device as truthful; unstressed samples that are classified as absent of any stress).

III-B-1. VSA Core Study: the IASCP Team CVSA Data

The detection of the presence of "blocking" (or nonblocking) on the charts was performed by the two CVSA operators that make up the IASCP team. (Recall that "blocking" occurs when the speaker produces speech under psychological stress; thus, blocking is predicted for both deceptive and stressed speech samples). The IASCP operators did so separately and without knowledge of either specific sample selection or the judgments made by the other operator. Their judgments were collated by a technician for subsequent presentation and statistical analysis.

To reiterate, the experiment was double-blind in nature and, as stated, it involved having each operator make the 300 forced-choice binary decisions privately. Their judgments were then processed by comparing them to the relevant stimuli (deception with and without jeopardy, high and low stress truth and so on). A large number of comparisons/analyses were carried out on the resulting data. Many of these can be found in the Technical Results section. However, a small number of the primary comparisons were grouped and sorted into 2x2 matrixes for use in this section (i.e., Summary Results). They permit perceived stress (as provided by CVSA analysis in terms of "blocking" and "no blocking") to be compared to actual stress -- and perceived deception and truth (also purportedly indicated by "blocking" and "no blocking," respectively, in CVSA) compared to actual deception and truth.

Consider the following matrix. It provides information about the identification of stress (only) in speech and voice by the IASCP team using the VSA (or core) database.

		Actual Condition: High Stress	Actual Condition: Low Stress
CVSA Analysis	High Stress *(Blocking)*	57% *(True Positive)*	62% *(False Positive)*
	Low Stress *(No Blocking)*	43% *(False Negative)*	38% *(True Negative)*

Figure 2: Identification of stress in speech samples from the VSA database (IASCP team).

As can be seen, this table provides a graphic view of any sensitivity displayed by the CVSA system to discriminate between speech produced under high stress conditions as well as those involving speech uttered under low stress conditions. Note that the identification of high stress falls above 50% (i.e., 57%) but that the false positive rate is even higher (62%). Further, conditions of low stress are not accurately identified. The relative similarity of the true positive and false positive rates is indicative of a lack of sensitivity to stress by the IASCP operators of the CVSA.

Perhaps more to the point are the data re: truth and deception. In this case, the contrast was between 1) the very low (stress) level where the statements were truthful and 2) deception produced under conditions of high jeopardy (frequently where the production of very offensive personal lies was combined with fear of electric shock).

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	64% *(True Positive)*	62% *(False Positive)*
	Truth *(No Blocking)*	36% *(False Negative)*	38% *(True Negative)*

Figure 3: Identification of deception and truth in speech samples from the VSA database (IASCP team).

The resulting data from the deception-truth utterances are of a class similar to those found for high-low stress statements. The detection of deception is higher (64%) but the judgments where truthful statements are judged to be falsehoods also remain high (62%). When the true positive rate (e.g. actual lies detected) is close or equal to the false positive rate (e.g. actual truths misclassified as deceptive) it is indicative of a device (and its operators) that is insensitive to the "signal" in question (in this case, deception). As would be expected, many variations of these comparisons will be found in the Technical Results section -- as will the statistical analyses. It will be seen there also, that the sensitivity analyses (i.e., d' or d prime) will provide the best perspective for understanding the relationships.

Consider next one of these variations. In this case the contrast is based on decisions made by multiple operators under conditions where 1) they agree on all judgments and 2) their confidence level is high. As can be seen in Figure 4, the rate at which deceptive samples are correctly classified is higher but, then, so is the false positive rate. Furthermore, the identification of truthful statements appears to suffer markedly.

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	75% *(True Positive)*	75% *(False Positive)*
	Truth *(No Blocking)*	25% *(False Negative)*	25% *(True Negative)*

Figure 4: Identification of deception and truth in speech samples from the VSA database (IASCP team – Operator Agreement Condition). These rates correspond to a subset of the samples – only those in which there was operator agreement on the presence/absence of blocking and only when those judgments were made with high confidence.

III-B-2. VSA Core Study: The NITV Team CVSA Data

As can be seen from Figures 5 and 6, the mean performance of the members of the NITV team was similar to that of the IASCP team. Certain patterns can be seen within the two figures. That is, the identification levels for speech uttered under stressful conditions, and when the utterances involve falsehoods spoken under high jeopardy exceed 50% (i.e., 61 and 65%, respectively). Taken alone, these data might suggest that the system (by itself) could be sensitive to stress or deception. Unfortunately, however, in both cases the false positive rates are even higher (70% in both instances). This relationship suggests that a high majority of low stress utterances -- **and** truthful speech -- would be classed as either high stress or deceptive. It should also be noted that the low stress truthful speech (30%) seen in Figure 6 would not be recognized as such in the great majority of instances. Further, Figure 5 can be compared to Figure 2, and Figure 6 to Figure 4. This comparison will tend to demonstrate that the IASCP team of two operators and NITV team of three, tended to perform similarly. The only differences (and they were but small ones) is that the NITV group appeared to be a little more aggressive in seeking deception and stress (note the 70% false positive rates).

		Actual Condition: High Stress	Actual Condition: Low Stress
CVSA Analysis	High Stress *(Blocking)*	61% *(True Positive)*	70% *(False Positive)*
	Low Stress *(No Blocking)*	39% *(False Negative)*	30% *(True Negative)*

Figure 5: Identification of stress in speech samples from the VSA database (NITV team).

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	65% *(True Positive)*	70% *(False Positive)*
	Truth *(No Blocking)*	35% *(False Negative)*	30% *(True Negative)*

Figure 6: Identification of deception and truth in speech samples from the VSA database (NITV team).

III-B-3. VSA Core Study: the Phonetician Team CVSA Data

As stated, the results in Figure 7 below are drawn from the responses of four phoneticians. The phonetician group was included to represent a group of untrained users to contrast with the IASCP team (i.e., a certified group of operators with limited experience) and the manufacturer's team (i.e., a certified group of operators with extensive experience representing the "best" operators possible). The phoneticians in question all have at least 25 years experience in decoding the complex wave forms of acoustic and related signals. Further, all have published in the area (several extensively) and are skilled in the interpretation of relevant data. They were asked, on a force choice basis, to identify those waveforms (i.e., from the two sets: 300 VSA; 56 SERE) that showed blocking and those that did not. They were provided instructions drawn from the NITV manual plus a number of examples. The same charts as those evaluated by the CVSA teams were presented to the phoneticians for analysis.

		Actual Condition: High Stress	Actual Condition: Low Stress
CVSA Analysis	High Stress *(Blocking)*	63% *(True Positive)*	62% *(False Positive)*
	Low Stress *(No Blocking)*	37% *(False Negative)*	38% *(True Negative)*

Figure 7: Identification of stress in speech samples from the VSA database (Phonetician team).

The resemblance between the data patterns produced by the two sets of CVSA operators and those of the phoneticians is striking. As can be seen from Figure 7, the

phoneticians identified but a modest number of the stressed speech signals (63%) as showing blocking, just as did the IASCP and NITV teams (57% and 61% respectively). All groups also logged a very high level of false positives (62%-70%) indicating that blocking occurred nearly two-thirds of the time no matter what the input stimulus. Moreover, the identification of speech uttered under conditions of low stress also ranged below chance (30-38%).

As can be seen in Figure 8, the patterns for blocking versus the deceptive and truthful utterances largely parallel those for stress. Moreover, the agreement of the phonetician's output with the other teams again is high. That is, it is a pattern of very modest identification of deception, poor identification of truth, and a high number of false positives. Overall, the phonetician operators of the CVSA showed no sensitivity to stress or deception in the VSA database, as indicated by the almost equal true positive and false positive rates.

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	65% *(True Positive)*	62% *(False Positive)*
	Truth *(No Blocking)*	35% *(False Negative)*	38% *(True Negative)*

Figure 8: Identification of deception and truth in speech samples from the VSA database (Phonetician team).

III-B-4 The SERE Field Study

This study was, of course, a much smaller one than that described above. Moreover, it was less controlled and, perhaps more important, was focused primarily on deception. That is, the targets were primarily untruths with decisions about stress and truthfulness relegated to a secondary role. Moreover, since all of the answers were one-word utterances (i.e., "yes" and "no") the SERE data base could only be used to evaluate the CVSA equipment. As can be seen from Figures 9 and 10, both the IASCP and NITV teams provided very similar judgments in their assessment of these materials.

Perhaps the most surprising thing about this analysis was that both teams showed a very low level of falsehood identification, with scores of 23% (IASCP team) and 19% (NITV team). False positive rates were also lower (41% and 45%) than those observed in judgment of the VSA database. These false positive rates are on average about double the "hit" rates for deception. The fact that these subjects were speaking during high stress when they lied could be confirmed by their very high heart rates. More informally, their demeanor also suggested high stress. Admittedly, observation of their physical behavior when being interrogated is hardly scientific. Nonetheless, their obvious discomfort with the session was consistent with the elevated heart rates.

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	23% *(True Positive)*	41% *(False Positive)*
	Truth *(No Blocking)*	77% *(False Negative)*	59% *(True Negative)*

Figure 9: Identification of deception and truth in speech samples from the SERE database (IASCP team).

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	19% *(True Positive)*	45% *(False Positive)*
	Truth *(No Blocking)*	81% *(False Negative)*	55% *(True Negative)*

Figure 10: Identification of deception and truth in speech samples from the SERE database (NITV team).

The phoneticians also participated in the field study. In this case, the phoneticians did a little better than the two other groups of operators in decoding the blocking—nonblocking configurations associated with deceptive speech and truthfulness (a deception-truth mean of 44% vs. means ranging 37-41% for the others). Nonetheless, and as can noted from observation of Figure 11, the outcome shows chance-level performance from all groups. (See also the sensitivity analyses reported in the Technical Results sections).

		Actual Condition: Deception	Actual Condition: Truth
CVSA Analysis	Deception *(Blocking)*	38% *(True Positive)*	49% *(False Positive)*
	Truth *(No Blocking)*	62% *(False Negative)*	51% *(True Negative)*

Figure 11: Identification of deception and truth in speech samples from the SERE database (Phonetician team).

III-B-5. The VSA Core Study: The IASCP Team LVA Data

The data found in Figure 12 are similar overall to most of those found for CVSA. The relationships identified by the IASCP team are not very encouraging. The rather low score (46%) in identifying high stress in speech is a case in point. Moreover, the false positive rate (60%) was quite high and comparable to the true positive rate, indicating a lack of sensitivity to stress.

		Actual Condition: High Stress	Actual Condition: Low Stress
LVA Analysis	High Stress	46% *(True Positive)*	60% *(False Positive)*
	Low Stress	54% *(False Negative)*	40% *(True Negative)*

Figure 12: Identification of stress in speech (IASCP team).

The values for deception-truth were not much better. They can be seen in Figure 13. Again the false positive rate (60%) was comparable to the true positive rate (50%), demonstrating a lack of sensitivity to deception.

		Actual Condition: Deception	Actual Condition: Truth
LVA Analysis	Deception	50% *(True Positive)*	60% *(False Positive)*
	Truth	50% *(False Negative)*	40% *(True Negative)*

Figure 13: Identification of deception and truth in speech (IASCP team).

III-B-6. The VSA Core Study: The V Team LVA Data

The results of the V team evaluation, shown in Figures 14 and 15, were comparable to those of the IASCP team in many respects. They showed slightly higher rates of identifying high stress speech (56% to 46% for IASCP) but were poorer in the low stress identifications (35% to 40% for IASCP) and for false positive errors (65% to 60% for IASCP). In any event, their true positive and false positive rates were similar enough to suggest that the LVA was not sensitive to deception in these speech samples.

		Actual Condition: High Stress	Actual Condition: Low Stress
LVA Analysis	High Stress	56% *(True Positive)*	65% *(False Positive)*
	Low Stress	44% *(False Negative)*	35% *(True Negative)*

Figure 14: Identification of stress in speech (V team).

As may be seen from Figure 15, the V team operators scored around chance when they attempted to correctly identify deception (52% for deception; 48% for incorrectly indicating that truthful statements are deceptive). They did have a somewhat lower false positive rate than often was seen in these types of data but, at 40%, it still is unacceptably high. Perhaps the most positive feature found in this analysis was that this team was able to correctly identify truthful statements, when they occurred, about 60% of the time.

This last evaluation completes the Summary Results section. Further interpretation of these data will be deferred until the Technical Results section to follow is complete. The presentation to follow will be both more varied and complete; the statistical evaluations also will be found there.

		Actual Condition: Deception	Actual Condition: Truth
LVA Analysis	Deception	52% *(True Positive)*	40% *(False Positive)*
	Truth	48% *(False Negative)*	60% *(True Negative)*

Figure 15: Identification of deception and truth in speech (V team).

III-C Technical Results
III-C-1 CVSA Testing
III-C-1-a CVSA Testing with VSA Database: IASCP Team
Table 1 provides the percentage of "blocking" responses collected from the IASCP team. (Recall that blocking refers to the gross shape of the signal displayed within a CVSA chart, namely a rectangular form; blocking is supposed to be the byproduct of psychological stress, which can be induced by a number of stressors, including lies produced with jeopardy). These percentages were averaged for two operators of the device, both of whom were certified by the manufacturer as competent in its use. Seven separate analyses were carried out; they are as follows:

- Stressed vs. unstressed materials (Analysis 1)
- Nondeceptive vs. deceptive materials (Analysis 2)
- Stressed vs. unstressed materials with deception absent (Analysis 3)
- Stressed vs. unstressed materials when deception was present (Analysis 4)
- Nondeceptive vs. deceptive materials when stress was low (Analysis 5)
- Nondeceptive vs. deceptive materials when stress was high (Analysis 6)
- Extreme groups design, in which only high-stress lies and low-stress truthful statements were examined (Analysis 7)

In all seven measures, the true positive rates were found to be near-chance (= 50%), ranging from 52% to 64%. Of course, true positive rates alone are not indicative of the sensitivity of the system to deception or stress. It is important to take into account the bias of the person or machine that is attempting to detect these speech attributes. For example, the examiner/machine may be predisposed to report that "deception is present" (i.e., is "liberal" in classifying a speech sample as deceptive), or the examiner may be biased toward reporting that "deception is absent" (i.e., is "conservative" in classifying a speech sample as deceptive). To eliminate these inclinations, true positive rates must be compared with false positive rates. The more alike the two proportions are, the less sensitive the device is to deception or stress. An examination of this team's false positive rates shows that they are highly similar to its true positive rates, ranging between 52% and 62%.

Analysis	Accurate		Inaccurate	
	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to Stress (All Conditions)	61%	47%	53%	39%
2. Sensitivity to Deception (All Conditions)	58%	45%	55%	42%
3. Sensitivity to Stress (Deception Absent)	57%	38%	62%	43%
4. Sensitivity to Stress (Deception Present)	64%	48%	52%	36%
5. Sensitivity to Deception (Low Stress)	52%	47%	53%	48%
6. Sensitivity to Deception (High Stress)	64%	43%	57%	36%
7. Extreme Groups (High-Stress Lie vs. Low-Stress Truth)	64%	38%	62%	36%

Table 1. This table presents the CVSA evaluations by the IASCP team using the VSA database. It shows the percentage of samples with blocking for all seven analyses of the dataset. The rates that correspond to accurate performance are "Hit" and "True negative." The rates that correspond to inaccurate performance are "False positive" and "False negative."

Two other types of analyses were also conducted. They included: 1) the conversion of the hit and false positive rates reported in Table 1 to d' (d-prime), a metric of true sensitivity and 2) repeated measures ANOVAs of the proportion of stress/deception responses for each type of sample. Repeated measures ANOVAs are commonly used in studies such as this; however, they often are only conducted on true positive rates. On the other hand, the problem of detecting the presence of deception or stress in speech is an example of the larger problem of stimulus or signal detection. As stated the detection of any phenomenon, such as deception, it is important to take into account the cited bias of the person or machine that is attempting to provide the data. In an extreme example, a VSA device might classify 90% or more of all speech samples presented as "deceptive." This process could be due to a human operator who wishes for the process to provide strong positive results and interprets most system output as "deceptive." In such a scenario, almost every utterance that actually involves deception would be correctly identified (a 90% true positive rate). At first glance, such results might appear to demonstrate that the deception detector works extremely well. However, it also would incorrectly classify nearly all of the truthful utterances as "deceptive," resulting in

a very high "false positive" rate. Accordingly, such a device could not be considered an accurate instrument in the detection of deception.

To reiterate, the determination of a team's or system's true sensitivity to the presence of stress, or deception, the "true positive rates" (the rate at which stressed or deceptive utterances are correctly classified) must be calibrated by the system's "false positive" rate (the rate at which truthful utterances are classified as deceptive). This calibration procedure forms a significant portion of Signal Detection Theory, which is commonly used in analyzing the type of data collected for this project (Macmillan and Creelman, 2005). If deception is taken as an example, the true sensitivity to its presence, d', ranges between 0 and 4+, with 0 referring to no sensitivity at all and 4 (and upwards) corresponding to very high sensitivity (associated with consistently classifying both deceptive utterances as deceptive and truthful utterances as truthful). For this analysis, d' measures were used to determine if CVSA could actually detect deception and stress. The conversion of values in Table 1 to d' is shown in Figure 16.

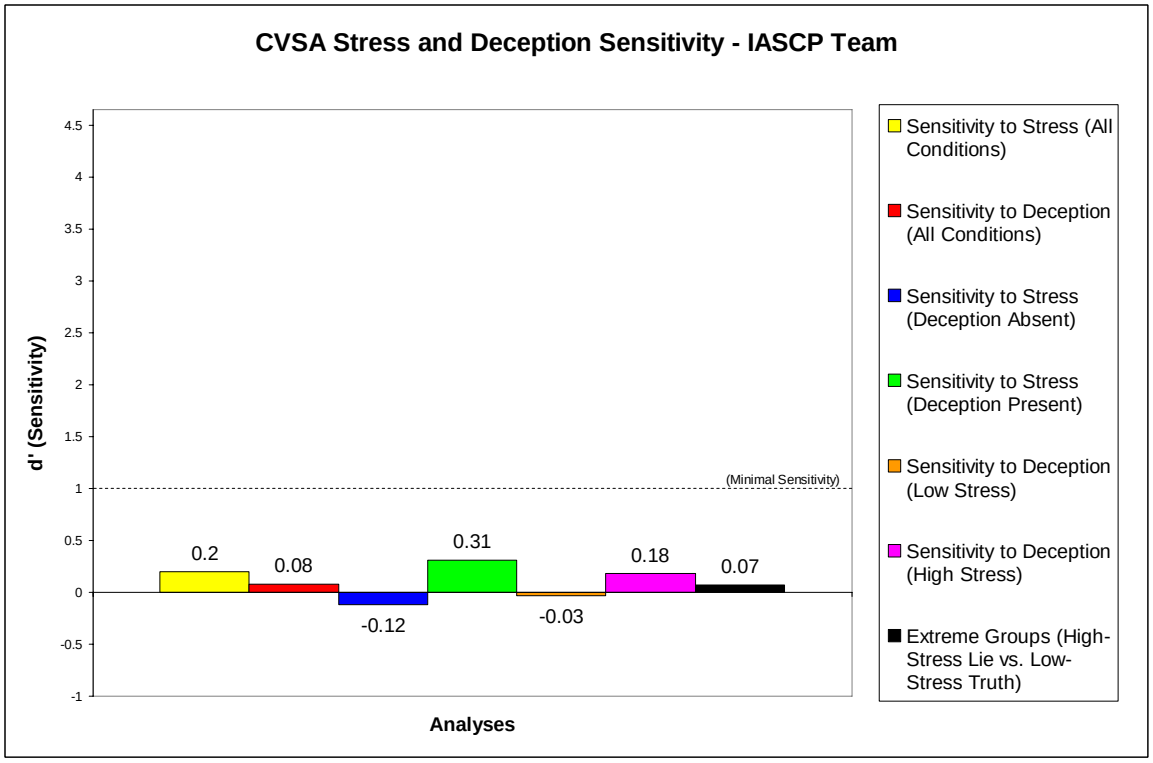

Figure 16: Sensitivity (d') measures for the IASCP team's operation of the CVSA using the VSA database. Seven different analyses are shown within this figure and are coded by color.

For a device and/or operator to be sensitive to a signal (deception and stress in this case), a d' value of four or higher would indicate excellent sensitivity. A value of one was set as the criterion corresponding to a *minimal* degree of sensitivity. Values that approximate zero indicate that the device and/or operator are not sensitive to stress/deception. Across all seven analyses, d' was low, ranging from -0.12 to 0.31.

Finally, a repeated measures ANOVA was conducted of the classification of the VSA database by the IASCP team, with Stress and Deception as within-subjects variables. Both variables, as well as their interaction proved to be nonsignficant, although the interaction did approach significance (Stress ($F(1,95) = 0.634$, $p = 0.43$; Deception ($F(1,95) = 0.08$, $p = 0.78$; Stress*Deception ($F(1,95) = 2.83$, $p = 0.10$). Post-hoc analyses were not conducted as none of the variables, or their interaction, were significant. The observed power for Stress, Deception and their interaction were all quite low (0.12, 0.06 and 0.38, respectively), indicative of large variability within the dataset.

III-C-1-b CVSA Testing with VSA Database: NITV Team

The responses from the NITV, averaged over three operators, are shown in Table 2. The same seven analyses with the IASCP team results appear here.

	Accurate		Inaccurate	
Analysis	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to Stress (All Conditions)	63%	39%	61%	37%
2. Sensitivity to Deception (All Conditions)	63%	39%	61%	37%
3. Sensitivity to Stress (Deception Absent)	61%	30%	70%	39%
4. Sensitivity to Stress (Deception Present)	65%	39%	61%	35%
5. Sensitivity to Deception (Low Stress)	61%	39%	61%	39%
6. Sensitivity to Deception (High Stress)	65%	39%	61%	35%
7. Extreme Groups (High-Stress Lie vs. Low-Stress Truth)	65%	30%	70%	35%

Table 2. This table presents the CVSA evaluations by the NITV team using the VSA database. It shows the percentage of samples with blocking for all seven analyses of the dataset. The rates that correspond to accurate performance are “Hit” and “True negative.” The rates that correspond to inaccurate performance are “False positive” and “False negative.”

When their performance was compared to that of the IASCP team, NITV operators showed a greater propensity to classify charts as “blocking.” This bias is illustrated in the slightly higher range of both true positive rates (61% - 65%) and false

positive rates (61% - 70%). The similarity of these two ranges suggests that the NITV operators and CVSA were not sensitive to either deception or stress. An examination of the corresponding d' values for these seven analyses confirms this observation (see Figure 17).

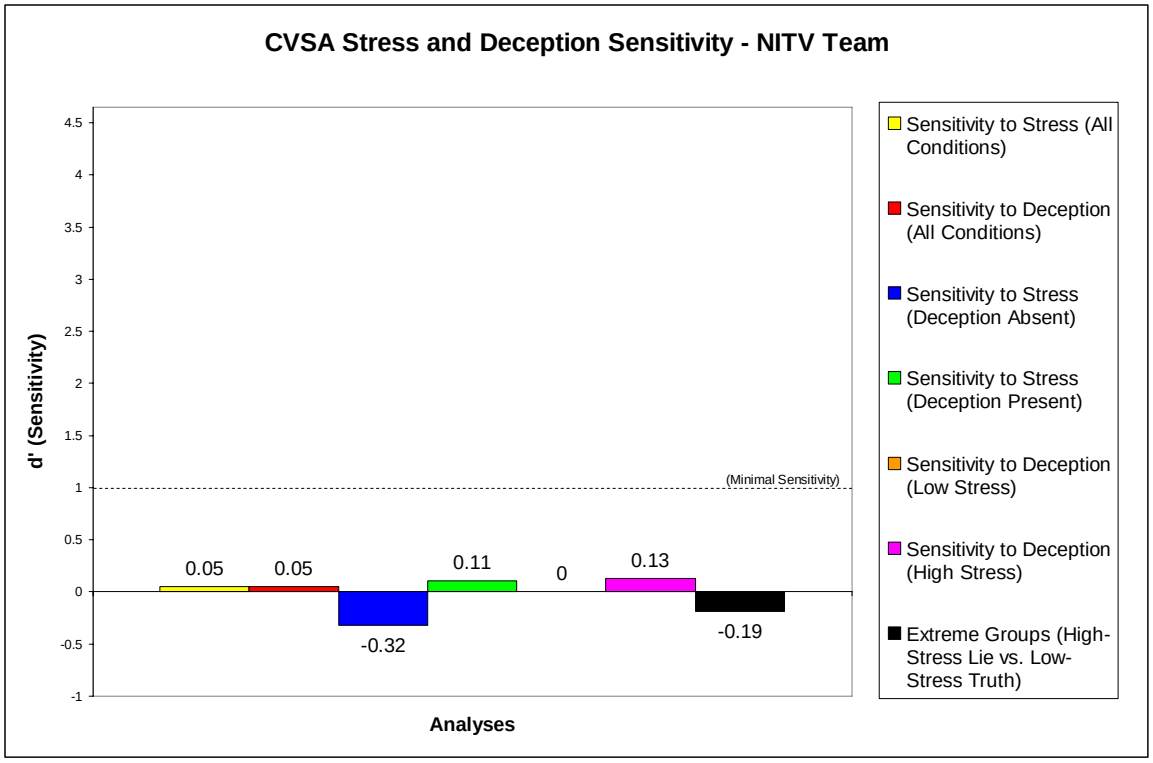

Figure 17: Sensitivity (d') measures for the NITV team's operation of the CVSA using the VSA database. Seven different analyses are shown within this figure and are coded by color.

In all analyses, d' values were very close to zero, ranging between -0.32 and 0.13. It should be recalled that values around zero correspond to no sensitivity on the part of an operator and/or device. The results were further examined in a repeated measures ANOVA, with Stress and Deception as within-subjects variables (as with the IASCP team data). Both variables, as well as their interaction proved to be nonsignficant, although the interaction again approached significance (Stress ($F(1,143) = 0.44$, $p = 0.51$; Deception ($F(1,143) = 0.33$, $p = 0.57$; Stress*Deception ($F(1,143) = 3.19$, $p = 0.08$). Post-hoc analyses were not conducted as none of the variables, or their interaction, were significant. The observed power for Stress, Deception and their interaction were low (0.10, 0.09 and 0.43, respectively), indicative of large variability within the dataset.

III-C-1-c CVSA Testing with VSA Database: Phonetician Team

The third team of operators to evaluate CVSA consisted of four phoneticians, specialists in the examination of speech signals for cues to the presence of different attributes of the signal (e.g., wave form, words, phonemes, stress, speaker gender, speaker age, talker identity). This team represented a group with minimal training in the

specific use of the device, although they represent decades of experience in general speech signal analysis. Their percentage of responses, organized by all seven analyses, appears in Table 3.

Analysis	Accurate		Inaccurate	
	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to Stress (All Conditions)	64%	46%	54%	36%
2. Sensitivity to Deception (All Conditions)	59%	43%	57%	41%
3. Sensitivity to Stress (Deception Absent)	63%	38%	62%	37%
4. Sensitivity to Stress (Deception Present)	65%	46%	54%	35%
5. Sensitivity to Deception (Low Stress)	54%	46%	54%	46%
6. Sensitivity to Deception (High Stress)	65%	37%	63%	35%
7. Extreme Groups (High-Stress Lie vs. Low-Stress Truth)	65%	38%	62%	35%

Table 3. This table presents the CVSA evaluations by the Phonetician team using the VSA database. It shows the percentage of samples with blocking for all seven analyses of the dataset. The rates that correspond to accurate performance are “Hit” and “True negative.” The rates that correspond to inaccurate performance are “False positive” and “False negative.”

The Phonetician team resembled both the IASCP and NITV teams in the relative similarity of their hit and false positive rates. Their true positive rates ranged between 54% and 65%; and such values fall near-chance for this task. As with the other teams, false positive rates were quite high, varying between 54% and 62%. Overall, the Phonetician team appeared no better or worse than either the IASCP team (who had received training) or the NITV team (a highly experienced group of operators). In addition, d’ values, shown in Figure 18, confirmed that the Phonetician team was not sensitive to stress or deception when using the CVSA with the VSA database.

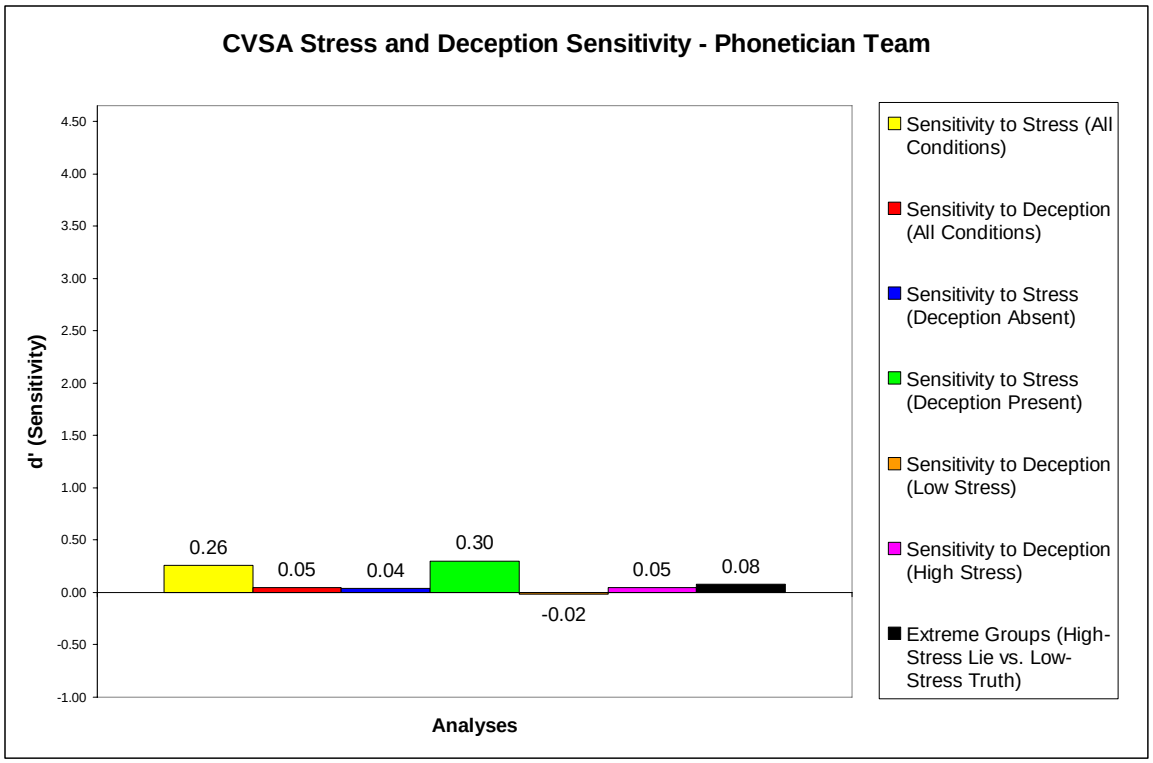

Figure 18: Sensitivity (d') measures for the Phonetician team's operation of the CVSA using the VSA database. Seven different analyses are shown within this figure and are coded by color.

Finally, the results were submitted to a repeated measures ANOVA, with Stress and Deception as within-subjects variables. Both variables, as well as their interaction proved to be nonsignficant, although the Stress variable approached significance (Stress ($F(1,191) = 3.01$, $p = 0.08$; Deception ($F(1,191) = 0.72$, $p = 0.40$; Stress*Deception ($F(1,191) = 2.52$, $p = 0.11$). Post-hoc analyses were not conducted as none of the variables, or their interaction, were significant. The observed power for Stress, Deception and their interaction were low (0.42, 0.14 and 0.35, respectively), indicative of large variability within the dataset.

III-C-1-d CVSA Testing with SERE Database: All Teams

The SERE database consisted of a smaller set of speech samples than did the VSA database, although ostensibly they constituted a more "natural" set of deceptive utterances produced under stress than those elicited in the laboratory. The present SERE materials, being only monosyllables, could only be used to evaluate the CVSA system. They were processed in two ways prior to CVSA input. First, the audio recordings were digitally extracted from the digital video files (sent to IASCP on individual CDs) and then segmented into individual audio files. Each file represents a single "yes" or "no" response by a SERE subject. Foils were also recorded to ensure that low-stress samples were included in the SERE database. They were added in order to fairly assess CVSA's sensitivity to psychological stress generated while lying.

Second, the foils and the original SERE samples were matched in background noise to ensure that external cues as to the nature of the speech materials being inputted did not become apparent. The SERE audiovideo recordings contained significant background noise, as is typical of materials recorded outside the highly controlled studio or laboratory environment. In contrast, the speech of the foil subjects was recorded in the Speech Perception Laboratory at the University of Florida under quiet conditions. To match the foil and SERE materials for background noise, a sample of the SERE noise was mixed with each foil file using signal processing software. The SERE database was then inputted to the CVSA computer using its sound card while following all the directions of the manufacturer.

Once all samples were inputted, the SERE materials were judged by all three teams. Only analyses for deception are shown in Table 4; that is this database did not consist of both stressed and deceptive samples in all combinations – only high-stress lies versus low-stress truth.

	Accurate		Inaccurate	
Team	True Positive	True Negative	False Positive	False Negative
IASCP	23%	59%	41%	77%
NITV	19%	55%	45%	81%
Phonetician	38%	51%	49%	62%

Table 4. This table presents the CVSA evaluations by all three teams (IASCP, NITV, Phonetician) using the SERE database. The rates that correspond to accurate performance are "True Positive" and "True negative." The rates that correspond to inaccurate performance are "False positive" and "False negative."

Interestingly, for this database, true positive rates were uniformly much lower across all teams than the false positive rates. However, true positive rates themselves were very low: only 19% - 38% of the lies were detected, with the least experienced team (the Phoneticians) showing the highest true positive rate. Of course, the Phonetician team also had the highest corresponding false positive rate, although it did not differ much from that of the NITV team.

While differences were seen in the comparison of the hit and false positive rates in these data, the conversion to d' failed to reveal that any team displayed even minimal sensitivity to deception in these materials (see Figure 19). All of the values were negative, as one would observe when true positive rates are actually below false positive rates.

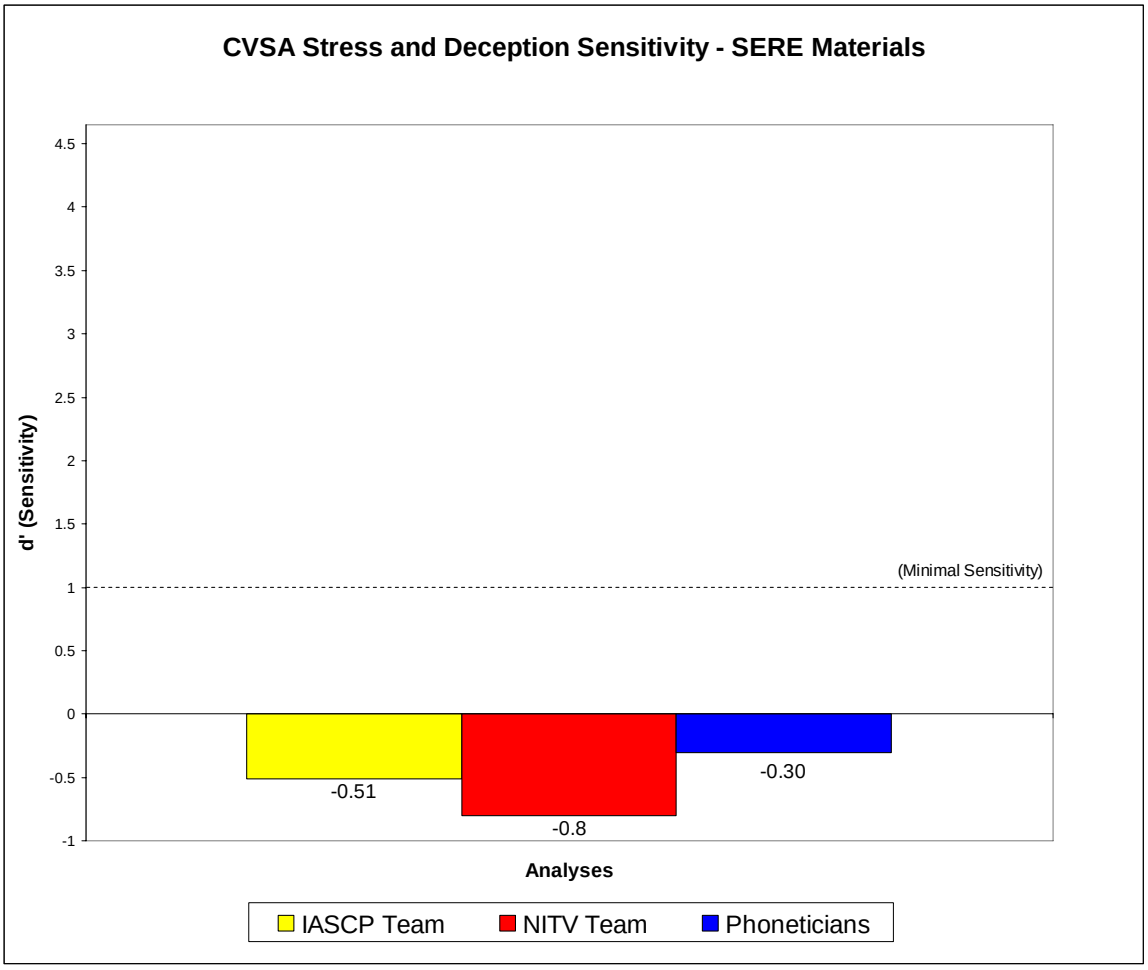

Figure 19: Sensitivity (d') measures for all three teams' operation of the CVSA using the SERE database.

The results of three repeated measures ANOVAs (one for each team) were consistent with the analyses reported above. For both the IASCP and Phonetician teams, no effect of the Deception variable was observed (IASCP: $F(1,47)=3.76$, $p = 0.06$; Phonetician: $F(1,95)=2.94$, $p = 0.09$). For the NITV team, a significant effect was observed ($F(1,71)=14.86$, $p < 0.01$), highlighting the large difference between the false positive rate (45%) and the true positive rate (19%). Because the false positive rate actually *exceeded* the true positive rate, this result meant that the NITV operators were significantly more likely to classify truthful SERE statements as deceptive as they were to correctly identify the deceptive SERE statements. Finally, it should be noted that all of the analyses conducted with the SERE-based field materials were made using a relatively small sample (56 speech tokens). This factor limited our ability to generalize from these findings.

III-C-1-e Interpretation of CVSA Testing

The CVSA did not display the expected sensitivity to the presence of deception, truth and/or stress in either the laboratory samples that constitute the VSA (core) database

or the smaller set of field materials (the SERE database). It should be stressed once again that the laboratory samples are the ones in which the presence of psychological stress during deception was verified through a range of measures (e.g., continuously recorded GSR and pulse rate; two self-report scales). In any event, the observed true positive and false positive rates varied by the particular team and by the particular analysis conducted. However, sensitivity, measured by d', only remained slightly above or below zero across all of these conditions. The conversion of the raw proportions to d' was critical in observing the performance of the equipment and its operators. Essentially, the d' analysis specifies CVSA's capacity to detect stress/deception (i.e., its true positive rate) by taking into account its tendency to also classify truthful and/or unstressed samples as deceptive and/or stressed.

While the raw data and all statistical analyses suggest only chance-level performance by the CVSA, alternate interpretations should be considered before classifying the device as wanting. For example, the present results with the VSA database could reflect limitations in the protocols used in its development. Essentially, the position could be taken that the stress shifts documented for the speech samples provided by the VSA database (i.e., those from the basic study) were not of a comparable magnitude to those induced in situations outside of the laboratory – i.e., those such as interrogations of individuals by police officers or military interrogators. In such cases, the "real-world" levels of stress might be higher than the psychological stress which can be generated in a laboratory setting on a college campus. University administrations carefully regulate the "use of human subjects" and place limits on how they can be treated in experiments. Indeed, this interpretation would be a difficult one to reject if only those speech samples that contained deception had been examined, i.e., if only true positive rates were assessed. However, an assessment of CVSA's performance on truthful and unstressed speech samples served as a robust control, one that permitted the examination of the device's potential bias to flag speech samples as deceptive in either the presence or absence of stress due to deception. If the VSA database, collected under highly-controlled conditions within the laboratory, contained inadequate levels of "real-world" stress, then false positive rates near zero would be expected. Such was not the case.

III-C-2 LVA Testing

All 300 samples from the VSA (core) database were transferred directly into the LVA software, following the manufacturer's instructions, along with their required calibration samples. That is, every sample from a given speaker (including low-stress truthful statements, high-stress truthful statements, low-stress deceptive statements and high-stress deceptive statements and soon) was paired with a general passage (drawn from the Rainbow passage) produced by that same speaker. The Rainbow passage served as calibration material for LVA. The reason for doing so is that software requires a sample of speech which is of sufficient duration to establish the norms for the speakers with respect to the voice parameters which LVA purportedly measures. After the VSA database was transferred into LVA, all of the sample statements (i.e., the speech material other than the Rainbow Passage) were marked as "Relevant." It should be noted that coding speech material as "Relevant" is a necessary step in the operation of LVA. Only the analysis of the "Relevant" speech materials is summarized in this report.

The LVA analysis itself was conducted differently by the two teams of evaluators, the IASCP team and the V team (e.g., two operators representing the manufacturer). The IASCP team at the University of Florida developed a protocol that did not require judgments by humans. This protocol was based on the training received by the two members of the team who are currently certified to use the device. The protocol varied depending on whether or not LVA was being operated to detect deception or stress. For truthful and deceptive samples, the “Final Analysis” in the "Show Report" menu in the Offline mode was examined. If the Final Analysis stated that "Deception was indicated in the relevant questions" for any appropriate segment, the neutral sentence (i.e., the relevant material) was coded as "deceptive.” (Note: A segment is a short portion of the speech material transferred into LVA. It automatically apportions a digital audio file into segments – a process that is largely, though not entirely, outside the user’s control). For examining LVA's ability to detect stress, LVA’s "JQ" parameter was used. The parameter is defined (by LVA) as one that measures emotional stress (not “physical” stress). In fact, of all of the parameters representing emotional or cognitive states, JQ appeared to be most appropriate for the speech materials collected. Following the threshold described in the software manual, a sample was coded as "stressed" if the mean JQ level across all relevant segments (weighted for the duration of each segment) was 35 or greater; otherwise the sample was coded as "unstressed." For both the deception as well as the stress analysis, trained LVA operators collated the results for submission to descriptive and statistical analysis.

The V team did not follow the same protocol as that developed by the IASCP team. Over the course of the study, the IASCP group was unable to reach agreement with V, LLC (the distributor of Nemesysco’s LVA software in the United States) on the analysis protocol reported here (see Appendix C for documents related to the relevant discussions with V). Ultimately, the V team conducted its own LVA test of the VSA database while at the University of Florida site. The V team did not use a consistent protocol with all samples and, therefore, no attempt to document their operation of the device can be made. However, the operators were both highly experienced users selected by the manufacturer. Thus, it can reasonably be expected that the V team’s use of the device was within the manufacturer’s guidelines.

III-C-2-a LVA Testing with VSA Database: IASCP Team

An analysis of LVA’s output for the relevant database was carried out by the IASCP team. In doing so, “true positive rates” were calculated for each sample type (e.g., low-stress lies, low-stress truths, etc.) as well as the “false positive,” “false negative” and “true negative” rates. Table 5 provides the percentage of responses of “Deception Indicated” as well as the percentage of samples that reached or exceeded the prescribed JQ threshold. The dataset was examined by means of seven related analyses:

- All stressed vs. unstressed materials (Analysis 1)
- All nondeceptive vs. deceptive materials (Analysis 2)
- Stressed vs. unstressed materials with deception absent (Analysis 3)
- Stressed vs. unstressed materials when deception was present (Analysis 4)
- Nondeceptive vs. deceptive materials when stress was low (Analysis 5)
- Nondeceptive vs. deceptive materials when stress was high (Analysis 6)

- By an extreme groups design, in which only high-stress lies and low-stress truthful statements were examined (Analysis 7)

Analysis	Accurate		Inaccurate	
	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to Stress (All Conditions)	48%	39%	61%	52%
2. Sensitivity to Deception (All Conditions)	47%	50%	50%	53%
3. Sensitivity to Stress (Deception Absent)	46%	40%	60%	54%
4. Sensitivity to Stress (Deception Present)	50%	37%	63%	50%
5. Sensitivity to Deception (Low Stress)	42%	46%	54%	58%
6. Sensitivity to Deception (High Stress)	46%	50%	50%	54%
7. Extreme Groups (High-Stress Lie vs. Low-Stress Truth)	50%	40%	60%	50%

Table 5. The percentage of samples coded as "stressed" or "deceptive" by LVA with the VSA database, employing the analysis developed by the IASCP team. It shows the percentage of samples with blocking for all seven analyses of the dataset. The rates that correspond to accurate performance are "True positive" and "True negative." The rates that correspond to inaccurate performance are "False positive" and "False negative."

In all seven measures, the true positive rates were below or near-chance (=50%), ranging from 42% to 50%. Moreover, an examination of the false positive rates shows that they are highly similar to the true positive rates (actually, they are slightly higher), ranging between 54% and 63%. Highly comparable true positive and false positive rates indicate a lack of sensitivity to the signal (in this case, "signal" refers to stress as well as deception).

The conversion of the true positive and false positive rates in Table 5 to the d' statistic reveals the same trend as was that observed in the CVSA dataset. Figure 20 provides the d' scores for the seven analyses. The seven d' values, ranging from -0.35 to -0.08, are well below the threshold for even a limited degree of sensitivity to deception or stress, let alone the threshold for being characterized as "accurate" or "sensitive."

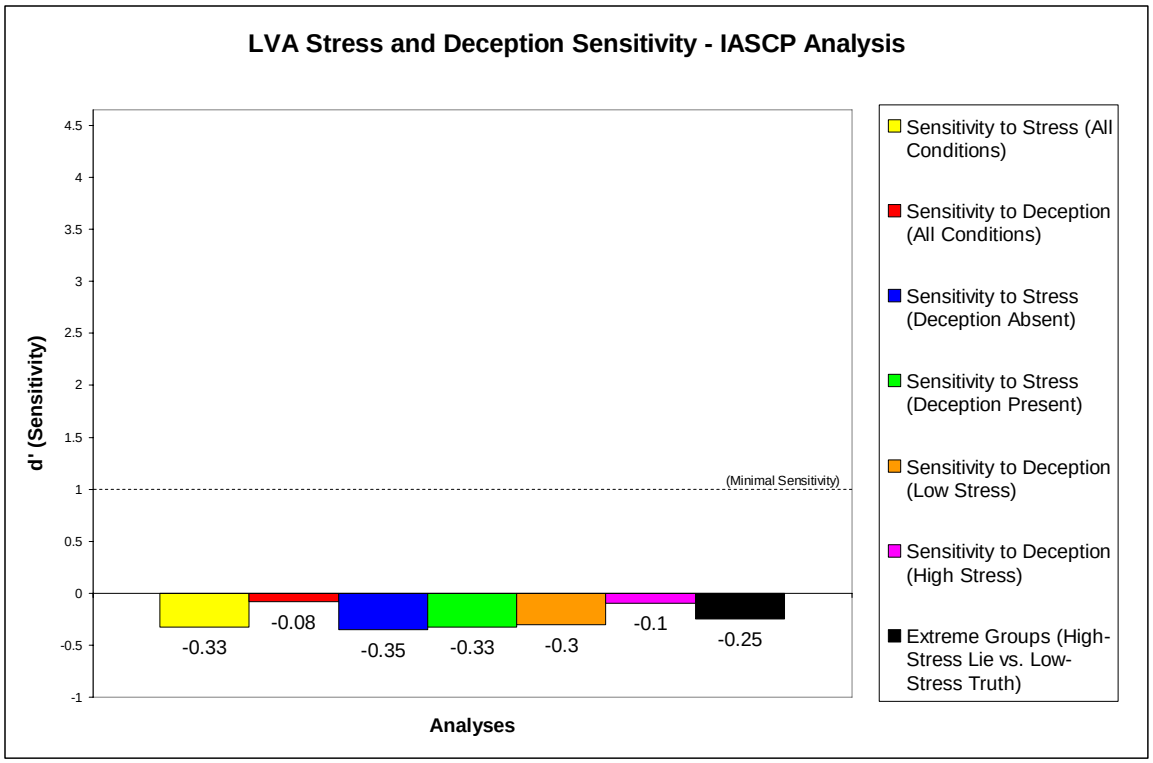

Figure 20: Sensitivity (d') measures for the IASCP team's analysis of the LVA using the VSA database. Seven different analyses are shown within this figure and are coded by color.

Two separate repeated measures ANOVAs also were conducted for evaluating LVA's performance with the VSA database from the basic study: One (the stress analysis) used the raw JQ values and the other (the deception analysis) used the "Deception Indicated" (DI) counts from the "Final Analysis" in the "Show Report" menu in the Offline mode. In the stress analysis, the unstressed and the stressed sample means were virtually identical (mean JQ = 36 and 34, respectively) and nonsignificant in difference ($F(1,95) = 2.98$, $p = 0.09$). For the truthful versus deceptive speech samples, the DI rates were not significantly different ($F(1,95) = 1.40$, $p = 0.24$). Further, observed power was low (Stress: 0.40; Deception: 0.22), indicating that LVA was inconsistent in correctly classifying unstressed, stressed, deceptive and nondeceptive materials.

III-C-2-b LVA Testing with VSA Database: V Team

The responses from the V team are provided by Table 6. These are *not* averaged values over the two operators. Instead, the V operators requested and were allowed to consult together and offer a single final judgment for each speech sample. Yet when the V team's results are examined, their true positive rates were in a similar range as those seen for the IASCP team's analysis (and all were close to chance). False positive rates were also quite high and exceeded the true positive rates in all but two analyses ("Sensitivity to Deception" and "Extreme Groups"). The conversion of these raw values to d' scores (see Figure 21) reveals the device's insensitivity to stress and deception in the VSA database, with values hovering near zero (-0.40 to 0.30). Two repeated measures

ANOVAs were also conducted, separately for stress and deceptive materials. Neither factor was significant (Stress: $F(1,94) = 1.79$, $p = 0.18$; Deception: $F(1,94) = 0.49$, $p = 0.49$) and the observed power was low (Stress: 0.26; Deception: 0.11).

Analysis	Accurate		Inaccurate	
	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to Stress (All Conditions)	56%	41%	59%	44%
2. Sensitivity to Deception (All Conditions)	47%	45%	55%	53%
3. Sensitivity to Stress (Deception Absent)	56%	35%	65%	44%
4. Sensitivity to Stress (Deception Present)	56%	36%	64%	44%
5. Sensitivity to Deception (Low Stress)	43%	41%	59%	57%
6. Sensitivity to Deception (High Stress)	52%	54%	46%	48%
7. Extreme Groups (High-Stress Lie vs. Low-Stress Truth)	52%	60%	40%	48%

Table 6. The percentage of samples coded as "stressed" or "deceptive" by LVA with the VSA database, as operated by the V team. It shows the percentage of samples with blocking for all seven analyses of the dataset. The rates that correspond to accurate performance are "True positive" and "True negative." The rates that correspond to inaccurate performance are "False positive" and "False negative."

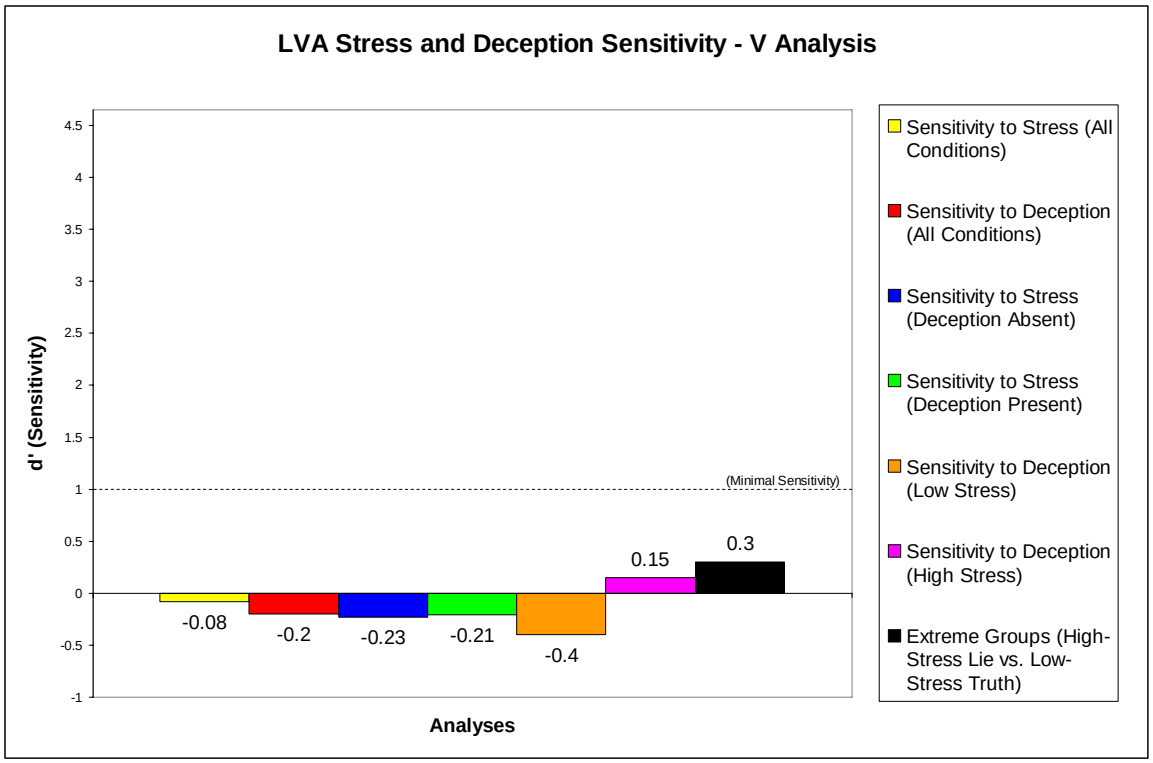

Figure 21: Sensitivity (d') measures for the V team's operation of the LVA using the VSA database. Seven different analyses are shown within this figure and are coded by color.

III-C-2-c Interpretation of LVA Testing

The performance of LVA on the VSA database by both the IASCP and V teams was similar to that observed with CVSA. That is, neither device showed significant sensitivity to the presence of stress or deception in the speech samples tested. The true positive and false positive rates were parallel to a great extent.

When discussing the CVSA results, we considered that they might have indicated a failure of our basic study protocol to elicit sufficiently stressed/deceptive speech samples due to the presumed inherent limitations of academic laboratory research. This interpretation was rejected due to the high false positive rates observed. This objection is even less plausible when the LVA system is considered. That is, its manufacturer claims that the device detects a wide variety of cognitive and emotional states. To do so, it must not only be sensitive to the relationship of the acoustic cues in the speech signal to the states in question, it also must exclude all other candidate cognitive states. CVSA's manufacturer, NITV, does not make as strong a claim about their product. Rather, in the training that the two IASCP team members received from NITV, the device is said to simply be a psychological stress detector. NITV advocates the use of standardized question format to ensure that the psychological stress detected by CVSA is due to actively lying, as opposed to some other stressor. Thus, CVSA purportedly detects a single cognitive state, and does not have the added burden of discriminating one cognitive state (e.g., stress due to deception) from other highly affective ones that LVA is claimed to detect (e.g., stress due to past traumatic experiences, degree of concentration,"

sexual arousal, imagination level, to name a few). For LVA to discriminate among a large set of cognitive states, it must be highly sensitive to whatever acoustic attributes of the speech signal cue those states. Presumed sensitivity at such levels suggests that LVA should be able to perform well with our laboratory samples as they contain both deception and documented levels of significant stress. However, unlike CVSA, LVA's false positive rates were consistently higher than their corresponding true positive rates. When both of these rates were converted to a single d', no actual sensitivity to stress and deception could be observed.

However, if it still is argued by LVA that the present laboratory protocols failed to elicit stress and deception that is sufficiently similar to stress and deception in a natural settings, the inclusion of unstressed speech samples and truthful speech samples in the database can address this concern. That is, if measurable stress/deception are not present in these samples, LVA should not have detected stress/deception in any portion of them. In fact, roughly half of the unstressed and truthful samples were classified by LVA as stress and deceptive, respectively. A device that is, in fact, sensitive to these states should not falsely detect them within the laboratory samples if we failed to elicit these qualities when using our protocol.

IV. SUMMARY AND CONCLUSIONS

The results reported in this study represent a complete evaluation of the devices using the speech material generated in the laboratory (the VSA database) and the field materials (the SERE database). That is, the tasked project was completed, as were several additions and upgrades. Specifically, a large database of highly controlled samples of spoken falsehoods and stressed speech was established, as well as a smaller one of field materials. The findings generated by this study led to the conclusion that neither the CVSA nor the LVA showed any sensitivity to the presence of deception or stress. Several analyses of subsets of the data were undertaken to explore any possibility that either system could perform under more controlled conditions, but no sensitivity was observed in any of these analyses (see the Technical Results section). These results are congruent with those observed in past surveys of research on voice stress analysis (National Research Council, 2003).

In discussing the results with the VSA database, we considered that they might have indicated a failure of our basic study protocol to elicit sufficiently stressed/deceptive speech samples due to the presumed inherent limitations of academic laboratory research. This interpretation was rejected due to the high false positive rate observed. More specifically, if the deceptive samples were produced without any real jeopardy, then neither device should have detected deception or stress in our *truthful* speech samples. Similarly, if we failed to elicit sufficient stress in what we classify as our high-stress speech samples collected in the laboratory, then neither device should have detected stress in our *low-stress* speech samples. In fact, both devices misclassified the low-stress and truthful samples with great frequency. Thus, these high rates of false positives cannot be explained by alluding to any inherent limitations of academic laboratory research. Moreover, the same pattern of results was observed with the SERE database, which consists of more naturalistic materials.

It should be noted that, while a large amount of research has been successfully completed, additional research is needed both to explore the basic relationships between speech and deception and to develop a richer database of speech samples for the evaluation of future commercial voice stress analyzers. These additional speech samples should be collected under simulated field conditions as well as "true field" samples, consisting of high jeopardy lies that are present and verifiable. The latter speech materials are the most difficult to acquire, but constitute the set was the greatest external validity.

The basic research program, developed under the current contract, should also be extended to the development of a *cross-language* voice stress analysis database (XL-VSA) paralleling the current VSA database in terms of the procedures used for the elicitation of stressed and unstressed truthful and deceptive speech samples. Commercial VSA products are increasingly sold in a global market. Security applications of the voice stress analysis type require systems that can handle a vast range of languages and dialects. This is especially true given the mobility of the world's peoples in a global economy and given the distribution of military assets in the Middle East, East Asia and many other regions. To date, no research has been conducted on the validity of any models of stress and deception in voice for speakers of different languages. This research should provide robust information – and databases – necessary to detect deception in the field.

Finally, this expanded (basic) research effort would provide a way to extract possible stress-truth-deception related parameters from the speech signal. In turn, the understanding of such vectors would provide manufactures with additional approaches which could be used in the design of more effective detection devices. Perhaps of greatest importance, such data would provide methods which, when combined with other types of behavioral assessments, could be potentially effective in the development of multiple-factor systems designed to reliably detect/identify the cited behaviors.

References

Abbs, J.H. and Gracco, V.L. (1984) Control of Complex Motor Gestures: Orofacial Muscle Responses to Load Perturbations of the Lip During Speech, *Journal of Neurophysiology*, 51:705-723.

Barland, G. (1975) Detection of Deception in Criminal Suspects, PhD dissertation, University of Utah, Salt Lake City.

Bassett, J.R., Marshall, P.M., Spillane R. (1987). The Physiological Measurement of Acute Stress (Public Speaking) in Bank Employees. *International Journal of Psychophysiology*, 5:265-73.

Bohnen, N., Nicolson, N., Sulon, J., Jolles, J. (1991) Coping Style, Trait Anxiety and Cortisol Reactivity During Mental Stress, *Journal of Psychosom Research*, 35:141-147.

Bossert, S., Berger, M., Krieg, J.C., Schreiber, W., Junker, M., Von Zerssen, D. (1988) Cortisol Response to Various Stressful Situations: Relationship to Personality Variables and Coping Styles, *Neuropsychobiology*, 20:36-42.

Brenner, M. and Branscomb, H.H. (1979) The Psychological Stress Evaluator, Technical Limitations Affecting Lie Detection, *Polygraph*, 8:127-132.

Brenner, M., Branscomb, H.H. and Schwartz, G.E. (1979) Psychological Stress Evaluator – Two Tests of a Vocal Measure, *Psychophysiology*, 16:351-357.

Brockway, B.F., Plummer, O.B. and Lowe, B.M. (1976) The Effects of Two Types of Nursing Reassurance Upon Patient Vocal Stress Levels as Measured by a New Tool, the PSE, *Nursing Research*, 25:440-446.

Cestaro, V.L. (1996) A Comparison of Accuracy Rates Between Detection of Deception Examinations Using the Polygraph and the Computer Voice Stress Analyzer in a Mock Crime Scenario. Report No. DoDPI95-R-0004. Ft. McClellan, AL: U.S. Department of Defense Polygraph Institute.

Cestaro, V.L. and Dollins, A.B. (1994) An Analysis of Voice Responses for the Detection of Deception. Report No. DoDPI94-R-0001. Ft. McClellan, AL: U.S. Department of Defense Polygraph Institute.

Chin, S.B. and Pisoni, D. (1997) *Alcohol and Speech*, San Diego, Academic Press.

Cummings, K. and Clements, M. (1990) Analysis of Glotal Waveforms Across Stress Styles, *Proceedings of the IEEE, ICASSP*, CH 2847:369-372.

Frankenhaeuser, M., Dunne, E., Lundberg, U. (1976) Sex Differences in Sympathetic-Adrenal Medullary Reactions Induced by Different Stressors, *Psychopharmacology*, 47:1-5.

Greaner, J. (1976) Validation of the PSE, unpublished MA thesis, Florida State University.

Haddad, D., Walter, S., Ratley, R. and Smith, M. (2002) Investigation and Evaluation of Voice Stress Analysis Technology, U.S. Dept. Justice, Report Grant 98-LB-VX-A103.

Heisse, J.W. (1976) Audio Stress Analysis – A Validation and Reliability Study of the Psychological Stress Evaluator (PSE), *Proceedings of the Carnahan Conference on Crime Countermeasures*, Lexington, KY, 5-18.

Hicks, J.W., and Hollien, H. (1981) The Reflection of Stress in Voice-1: Understanding the Basic Correlates, *Proceedings of the Carnahan Conference on Crime Countermeasures*, Lexington, KY, 189-194.

Hollien, H. (1980) Vocal Indicators of Psychological Stress, *Forensic Psychology and Psychiatry*, New York, New York Academy of Sciences, 47-72.

Hollien, H., Geison, L.L. and Hicks, J.W., Jr. (1987) Data on Psychological Stress Evaluators and Voice Lie Detection, *Journal of Forensic Sciences*, 32:405-418.

Hollien, H. (1990) *Acoustics of Crime*, New York, Plenum.

Hollien, H., Saletto, J.A. and Miller, S.K. (1993) Psychological Stress in Voice: A New Approach, *Studia Phonetica Posnaniensia*, 4:5-17.

Hollien, H., DeJong, G. and Martin, C.A. (1998) Production of Intoxication States by Actors: Perception by Lay Listeners, *Journal of Forensic Sciences*, 43:1163-1172.

Hollien, H., (2000) *Forensic Voice Identification*, London, Academic Press.

Hollien, H., DeJong, G., Martin, C.A., Schwartz, R. and Liljegren, K. (2001) *Journal of the Acoustical Society of Americ*a, 110:3198-3206.

Hollien, H., Liljegren, K. and Martin, C.A. (2001) Production of Intoxication States by Actors: Acoustic and Temporal Characteristics, *Journal of Forensic Sciences*, 46:68-73.

Hollien, H., and Schwartz, R., (2001) Speaker Identification Utilizing Noncontemporary Speech, *Journal of Forensic Sciences*. 46:63-67.

Horvath, F. (1978) An Experimental Comparison of the Psychological Stress Evaluator and the Galvanic Skin Response in Detection of Deception, *Journal of Applied Psychology*, 63: 338-344.

Horvath, F. (1979) The Effects of Differential Motivation on Detection of Deception with the Psychological Stress Evaluator and the Galvanic Skin Response. *Journal of Applied Psychology*, 64:323-330.

Horvath, F. (1982) Detecting Deception: the Promise and the Reality of Voice Stress Analysis, *Journal of Forensic Sciences*, 27:340-351.

Inbar, G.F. and Eden, G. (1976) Psychological Stress Evaluators: EMG Correlations with Voice Tremor, *Biolog. Cybernetics*, 24:165-167.

Janniro, M.J., and Cestaro, V.L. (1996) Effectiveness of Detection of Deception Examinations Using the Computer Voice Stress Analyzer. Report No. DoDPI96-R-0005. Ft. McClellan, AL: U.S. Department of Defense Polygraph Institute.

Kirschbaum, C., Wust, S., Hellhammer, D. (1992) Consistent Sex Differences in Cortisol Responses to Psychological Stress, *Psychosom Medicine*, 54:648-57.

Kubis, J. (1973) Comparison of Voice Analysis and Polygraph as Lie Detection Procedures, *Technical Report LWL-CR-U3B70*, Aberdeen Proving Ground, MD, U.S. Army Land Warfare Laboratory.

Kuenzel, H., (1994) On the Problem of Speaker Identification by Victims and Witnesses, *Forensic Linguistics*, 1:45-58.

Leith, W.R., Timmons, J.L. and Sugarman, M.D. (1983) The Use of the Psychological Stress Evaluator with Stutterers, *Journal of Fluency Disorders*, 8:207-213.

Lykken, D. (1981) Tremor in the Blood, New York, McGraw-Hill.

Lynch, B.E. and Henry, D.R. (1979) A Validity Study of the Psychological Stress Evaluator, *Canadian Journal of Behavioral Sciences,* 11:89-94.

Macmillan, N.A. and Creelman, C.D. (2005) Detection: Theory: A User's Guide Second Edition, New Jersey, Lawrence Erlbaum Associates.

Maier, W., Buller, R., Phillip, M. and Heuser, J. (1988) The Hamilton Anxiety Scale: Reliability, Validity and Sensitivity to Changing Anxiety and Depressive Disorders, *Journal of Defective Disorders,* 14:61-68.

Meyerhoff, J.L., Saviolakis, G.A., Koenig, M.L., and Yourick, D.L. (2000) Physiological and Biochemical Measures of Stress Compared to Voice Stress Analysis Using the Computer Voice Stress Analyzer (CVSA). Report No. DoDPI98-P-0004. Ft. Jackson, SC: U.S. Department of Defense Polygraph Institute.

McGlone, R.E. (1975) Tests of the Psychological Stress Evaluator (PSE) as a Lie and Stress Detector, *Proceedings of the Carnahan Conference on Crime Countermeasures*, Lexington, KY, 83-86.

McGlone, R.E., Petrie, C. and Frye, J. (1974) Acoustic Analysis of Low-Risk Lies, *Journal of the Acoustical Society of America*, 55:S20(A).

Nachshon, I. and Feldman, B. (1980) Vocal Indices of Psychological Stress: A Validation Study of the Psychological Stress Evaluator, *Journal of Police Science Administration*, 3:40-52.

National Research Council (2003) *The Polygraph and Lie Detection. Committee to review the scientific evidence on the Polygraph*, Division of Behavioral and Social Sciences and Education, Washington, DC, The National Academies Press.

Nejtek, V.A. (2002) High and Low Emotion Events Influence Emotional Stress Perceptions and are Associated with Salivary Cortisol Response Changes in a Consecutive Stress Paradigm, *Psychoneuroendocrinology*, 27:337-52.

Netsell. R. (1983) Speech Motor Control: Theoretical Issues with Clinical Impact, *Clinical Dysarthria,* San Diego, College Hill Press, 1-19.

Nolan, J.F., (1983) *The Phonetic Basis of Speaker Recognition*, Cambridge, UK, University Press.

O'Hair, D., Cody, M. J. and Behnke, R. R. (1985) Communication Apprehension and Vocal Stress as Indices of Deception, *Western Journal of Speech Communication*, 49:286-300.

Pisoni, D. and Martin, C.S. (1989) Effects of Alcohol on the Acoustic-phonetic Properties of Speech: Perceptual and Acoustic Analyses, *Alcoholism: Clinical Exper. Res.*, 13:577-587.

Scherer, K.R. (1981) Vocal Indicators of Stress, *Speech Evaluation in Psychiatry,* New York, Grune and Stratton, 171-187.

Scherer, K.R (1986) Voice, Stress and Emotion, *Dynamics of Stress: Physiological, Psychological and Social Perspectives*, New York, Plenum Press, 157-179.

Shipp, T. and Izdebski, K. (1981) Current Evidence for the Existence of Laryngeal Macrotremor and Microtremor, *Journal of Forensic Sciences*, 26:501-505.

Smyth, J., Ockenfels, M..C., Porter L., Kirschbaum, C., Hellhammer, D.H. and Stone, A.A. (1998) Stressors and Mood Measured on a Momentary Basis are Associated with Salivary Cortisol Secretion, *Psychoneuroendocrinology*, 23:353-70.

Stevens, K.N., (1971) Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds, *Proceedings of the Seventh International Cons. Phonetic Sciences*, Montreal, 206-232.

Van der Car, D.H., Greaner, J., Hibler, N., Speelberger, C.D. and Bloch, S. (1980) A Description and Analysis of the Operation and Validity of the Psychological Stress Evaluator, *Journal of Forensic Sciences*, 25:174-188.

Williams, C.E. and Stevens, K.N. (1972) Emotions and Speech: Some Acoustical Correlates, *Journal of the Acoustical Society of America*, 2:1238-1250.

Appendix A. Anxiety/stress tests

A.1 "Emotion Felt" checklist

RESPONSE TEST

Name/Code: _____ Date: _____

Type of Trial: _____ Experimenter: _____

Please rank how you felt with respect to the following five emotions. Did you feel:

Discomfort	**Stressed**	**Angry**	**Embarrassed**	**Anxious**
10 ☐	10 ☐	10 ☐	10 ☐	10 ☐
9 ☐	9 ☐	9 ☐	9 ☐	9 ☐
8 ☐	8 ☐	8 ☐	8 ☐	8 ☐
7 ☐	7 ☐	7 ☐	7 ☐	7 ☐
6 ☐	6 ☐	6 ☐	6 ☐	6 ☐
5 ☐	5 ☐	5 ☐	5 ☐	5 ☐
4 ☐	4 ☐	4 ☐	4 ☐	4 ☐
3 ☐	3 ☐	3 ☐	3 ☐	3 ☐
2 ☐	2 ☐	2 ☐	2 ☐	2 ☐
1 ☐	1 ☐	1 ☐	1 ☐	1 ☐
0 ☐	0 ☐	0 ☐	0 ☐	0 ☐

10 equals greatest intensity
1 equals least intensity
0 means that the emotion was not present

Please check the level you felt.

Thank you.

A.2 Modified Hamilton checklist

(Modified Hamilton)

Name/Code: _____ Date: _____

Type of Trial: _____ Experimenter: _____

WHAT DO YOU FEEL RIGHT NOW?

Symptom	Scale				
Sweating	1	2	3	4	5
Shaking	1	2	3	4	5
Shortness of Breath	1	2	3	4	5
Chest Pain	1	2	3	4	5
Nausea	1	2	3	4	5
Dizziness	1	2	3	4	5
Irritability	1	2	3	4	5
Distractibility	1	2	3	4	5
Muscle Tension	1	2	3	4	5
Irregular Heartbeat (Palpitations)	1	2	3	4	5

Directions: Please circle the number that corresponds to the symptoms you are experiencing right now. The scale is as follows:

1 – Absent
2 – Mild
3 – Moderate
4 – Severe
5 – Extreme

Score Total: _____ (out of 50)

Appendix B. Subjects whose speech materials were included in the Voice Stress Analysis database.

B.1 Male subjects

Subject Code	Mean GSR + Pulse	Mean Self-Report Scales	Overall Stress Shift
M104	29%	93%	61%
M105	105%	32%	68%
M106	35%	189%	112%
M110	59%	137%	98%
M111	54%	254%	154%
M112	229%	66%	147%
M113	146%	51%	99%
M114	112%	202%	157%
M116	93%	279%	186%
M117	72%	312%	192%
M118	38%	378%	208%
M119	70%	151%	110%
M120	60%	186%	123%
M122	24%	352%	188%
M123	95%	204%	149%
M124	76%	142%	109%
M125	104%	211%	157%
M127	69%	146%	108%
M129	55%	131%	93%
M130	70%	198%	134%
M131	29%	185%	107%
M134	83%	306%	194%
M135	63%	86%	74%
M138	10%	145%	78%

B.2 Female subjects

Subject Code	Mean GSR + Pulse	Mean Self-Report Scales	Overall Stress Shift
F201	22%	69%	45%
F202	26%	182%	104%
F205	67%	33%	50%
F208	32%	82%	57%
F213	115%	59%	87%
F214	24%	295%	160%
F215	40%	150%	95%
F216	61%	83%	72%
F217	385%	361%	373%
F218	65%	61%	63%
F219	467%	317%	392%
F220	99%	169%	134%
F221	108%	295%	202%
F223	94%	477%	286%
F224	95%	152%	123%
F227	66%	358%	212%
F229	134%	226%	180%
F230	70%	108%	89%
F231	55%	106%	80%
F232	23%	351%	187%
F233	196%	148%	172%
F235	76%	196%	136%
F236	165%	100%	133%
F238	129%	310%	220%

Appendix C. Correspondence between the IASCP team at the University of Florida and V, LLC, distributors of Nemesysco's LVA product, concerning the protocol to test LVA.

C.1 Email from IASCP to John Taylor of V, 11/10/05

John,

We've got the database all set here to test LVA. Since V has not responded to our request for a protocol, we thought we would present how we think the software should be used to test our materials based on the training Kevin Hollien and I received and based on the software's documentation.

As you recall, our database consists of passages read under various conditions of stress and deception. You received a demo of our data collection method. From those passages we recorded, we are taking:

1. Truthful, low stress samples
2. Truthful, high stress samples
3. Deceptive, low stress samples
4. Deceptive, high stress samples

Remember that these are the carrier phrases from our passages: linguistically neutral sentences that do not relate to the specific topic being discussed.

For each speaker that thas been recorded, we are pairing these carrier phrases with a longer, neutral passage. The longer neutral passage will serve as calibration for LVA and the carrier phrase will be marked as relevant.

Each of these samples will be analyzed in two ways. For truthful and deceptive samples, we will check the Final Analysis in the "Show Report" menu in the Offline mode. If the Final Analysis indicates that "Deception was indicated in the relevant questions" for any relevant segment, we will code that carrier phrase as "deceptive" according to LVA.

Our second analysis will test LVA's ability to detect stress induced by our laboratory procedures. we will use the "JQ" parameter for this purpose. If the average JQ level across all relevant segments (weighted for the duration of each segment) is 35 or greater, we will code that carrier phrase as "stressed"; otherwise "unstressed."

We would appreciate a response to this protocol. If you want to revise this protocol or substitute a different procedure, please describe your proposed changes in sufficient detail.

We look forward to hearing from you,

Jimmy Harnsberger

C.2 Response from V to IASCP, 11/17/05

Dr. Harry Hollien
University of Florida
50 Dauer Hall
Gainesville, FL 32611

Re: Layered Voice Analysis

Dr. Hollien

Thank you for providing us the opportunity to comment on the study protocols for Phase 1. We welcome a rigorous evaluation of Layered Voice Analysis ("LVA") technology, and are committed to working with you and your team to develop a full and complete understanding of the capabilities of LVA.

As we have discussed with you, we have concerns that the Phase 1 protocol may not provide the necessary sampling to measure deception. This is based on our continued skepticism about the methods and protocols used to collect the sample statements.

LVA is designed to detect deception based upon identifying an individual's intent to deceive. Based upon our understanding of the protocol, and after discussion with the developer and other scientists familiar with LVA, we question whether the voice samples to be used in this Phase reflect a true intent to deceive as measured by LVA.

As you know, the results of a previous study have been subject to extensive criticism because of the use of artificial attempts to create an equivalent to real-life deception. As we have stressed from the first meeting with DOD-CIFA, we believe that the analysis of voice samples of individuals in real life situations will provide the most accurate test of LVA's ability to detect deception and other emotional/psychological states of the speaker.

Despite these challenges and in the interest of the science and this specific research, V has decided to move forward with the analysis of the samples. We hope you will appreciate our position as well as our willingness to move forward in an effort to validate the quality of the samples and to further validate LVA's utility.

As we move forward on Phase 1 of the study, we will attempt to identify what LVA components can be used to best measure the subjects' states under the conditions you created in your laboratory. Based upon the little information we now have, given our inability to preview any data segments, the following is what we can tell you at this time.

There likely may be a great deal of variance in the results of the readings in this test from one session (the three readings by one reader) to the next. The issues that will affect the end result of each session will include the following:

- Strength of conviction on a subject for the individual,
- Wording of contradictive statement,
- Investment of reader to convince the actual in-lab listener(s) and potential listeners (as briefed to subject pre-test),
- Length of statement being read and the number of issues being covered (the longer the statement and the more the issues the greater the potential for the reader "to lose interest" or become distracted from the task),
- Whether there is a consistent introduction for calibration purposes, and
- The reader's particular relationship to an issue or personal experience with an issue (e.g., one reader may be pro-issue for specific reasons but not in general, or a reader may have had an unique issue related personal experience (i.e. abortion, drug use, killing someone as part of a military action) and consciously/sub-consciously regret the action, which could produce a variety of unpredictable emotional/psychological dynamics therefore creating unpredictable results in the LVA under the condition of your sample protocol).

The analysis may show that the act of reading itself takes cognitive effort and may shift the attention from the actual mental "veracity" assessment. One may even have a dual thought path when reading as if on "auto pilot". For the experiment to be truly controlled, this issue must be properly dealt with, and kept from happening. That is one of the reasons we generally do not favor the use of reading in LVA examinations.

The ratio between true statements and false statements must be taken into account as well. It is our understanding that most people do not lie continuously. In a normal conversation, people generally lie only when they really need too, and not upon request or in any constant manner. Even in cases where the whole investigation is around a particular deceptive statement, many non-deceptive statements will (or should) be expressed. Those are helpful to LVA's calculation of the baseline calibration of the conversation.

To the extent the already collected data sets will allow, we recommend the following:

- Do not analyze more than one deceptive/"laboratory created" conviction in an interview;
- If a reading session is lengthy, analysis should be weighted to earlier portions of session when the reader is most likely to be more engaged; and
- Look at the information on an issue in a larger context, not just based on one or two words.

Because we have no experience with the effects of shocking, and/or threatening to shock, people during statements, we cannot offer any definitive opinion as to the probable impact that will have on the LVA assessment.

Due to the reading and the overall circumstances of the session, there may be high JQ and stress throughout, and likely increasing on words the reader finds complicated words (this

would be typical of a news anchor reading from a teleprompter). This may at times create random deception pop-up messages related to the complicated words.

As, in our opinion, the subjects will be "acting-out" their attempted deception, SPJ may be inconsistent from subject to subject. Depending upon the subject's internal reaction to the scenario, it might either increase or decrease dramatically. We may see many "voice manipulation" pop-up messages instead of deception messages in the areas where the subject are "lying".

LJ and AVJ may be higher then the normal in any case (we would assume above 6 and 3 respectively) – but if the subject loses concentration and gets into an "auto pilot" state, it may drop down significantly. If the subject is mentally "fighting" with the statements or trying to further process his beliefs, the AVJ/LJ level may increase.

Fmain, SubCog, SubEmo activities may be unpredictable because they will be processing too many artificial variables. High ANT (Anticipation) should be found sporadically, but mostly in the "lies" area.

Based upon the limited information we have, our training faculty has identified the following possible outcomes for the Phase 1 protocol:

1. We suspect that there may be an elevated Stress (JQ) response for most individuals (around 25-25).
2. We suspect that there may an increase in parameter readings such as JQ, Anticipation, and Global Stress (AROUND 130+) just before the first shocking and that it may not present itself after that if the shock is of no physical or emotional significance.
3. We suspect that in general there will be an increase in global stress response readings (around 120-140).
4. We suspect that the SPT reading will may be elevated for males in the 300 range, females in the 400 range.
5. We suspect if the individual is not a strong reader, SPJ scores will be elevated (high end of normal 300).
6. We suspect most AVJ scores will be in the 3.5 to 6.0 range
7. We suspect that some persons (about 40%) will show abnormal scores in the imagination readings.
8. We suspect that deception will be found for those individuals that are truly lying about their convictions, but many or most readers may not intend or conceive of the contrary statements as "lying".
9. We suspect that there will be an occasional high SOS.
10. We suspect that the sub EMO will have more activity and will average between 15-30 depending on the issue.
11. We suspect that the sub Cog will have lesser activity depending on "prep" information (ranges 5-20) and how much extra cognition goes into the reading process.
12. We suspect we may see both a rise in stress prior to the shock and an immediate sharp increase of SPT/Emotional. A sharp decrease of any Cog. related parameters right after

the shock (SPJ, JQ, LJ) may result if the subject experiences an anger response to the pain.
13. We suspect that the complexity of the statements, as well as the reward / punishment concept and details (if applicable) may also have a material effect on the results (we do not have enough information to state more).

We look forward to working with you on this study.

Best regards,

C. David Watson
General Counsel, Chief Operating Officer

CC: Richard D. Parton, Ph.D.
John Taylor

C3. Reply to V, 12/10/05

C. David Watson
General Counsel, Chief Operating Officer
Layered Voice Analysis

Mr. Watson,

Thank you for providing an initial draft of protocols you feel useful for the testing of Layered Voice Analysis. In your comments preceding these protocols, you raised an issue about the methods we used to elicit samples for testing "voice stress analysis" software, including LVA. Specifically, you suggested that we did not verify that the samples were produced with a "true intent to deceive." We respectfully respond that the "intent" of speakers is information unavailable to anyone attempting to evaluate LVA -- or, for that matter, by anyone for any purpose whatsoever. "Intent" refers only to the speaker's motivations to produce the speech sample and the thoughts/emotions/cognitive state of speaker during an utterance. Currently, no technology exists which is capable of "reading people's minds" during any motor speech -- or any other -- activity. For example, even brain imaging technologies cannot be used to classify blood flow patterns into such specific "intents" as LVA purports to detect. And even to the limited extent that brain imaging technology can be used to observe cognitive states, it can only do so under extremely constrained laboratory conditions -- and not at all in the "real world" situations you cite as being the "most accurate test of LVA's ability to detect deception." Given the conflicting constraints you have suggested for a "fair" test of LVA (i.e., knowing the speaker's intent while that individual produces lies in a real-world situation), it appears impossible to develop any procedure at all that could "test" LVA. In fact, by your own admissions, it would appear impossible to determine the validity of your system on any level.

However, given your willingness to continue collaborating with us, we must assume that you predict that LVA will show some degree of sensitivity to deception when our speech samples (i.e., even those from our laboratory study) are processed by your system. Therefore, we are responding to your specific recommendations for testing LVA's sensitivity to deception. We are also responding to your specific recommendations for the testing LVA's sensitivity to psychological stress because our protocol ensured that we can objectively demonstrate that our subjects experienced a significant degree of that behavior (i.e., stress). We did so using procedures that were based on our measuring of physiological correlates of stress (as well as by self-reports and expert ratings).

Most of our responses to your comments will involve a request for greater detail relative to this first (i.e., basic) study of several we plan: We need to know 1) what parameter(s) you wish to have examined, 2) the thresholds (if applicable) we need to use for two general kinds of speech samples:
1. Samples in which the "signal" is present (signal refers to the speaker's deception and stress in the speech utterances).

2. Samples in which the signal is absent (these materials would include the truthful statements).

First, for your bulleted points on page 2, you list several issues that could affect our results. We must point out, however, that you did not appear, in many instances, to take our research procedures into account. We are only testing for high and low stress lies and for speech uttered under high and low stress. Indeed, we have employed extremely rigorous procedures in order to obtain the samples we use. As a matter of fact, we were only able to use (in this particular investigation) about half of the volunteers that attempted to meet our standards. Best yet, we were able to independently verify that the subjects we did include actually achieved the high stress (and low stress also) conditions that we sought. Finally, many of the parameters you refer to do not appear to have any relevance to our research as we do not intend to employ them. Nevertheless, we are willing to address them in this letter. Our responses are as follows:

1. “Strength of conviction.” We find this to be an irrelevant issue since it is impossible for anyone to actually quantify exactly how strongly someone feels about an issue. We recruited subjects who self-identified themselves as holding very strong convictions re: the issues they lied about. No independent means exist to measure a person’s strength of conviction. Hence, we all must rely on the subject’s self-report plus their behavior during the experimental trial (which was consistent with their self-reports for all subjects used).

2. “Wording of contradictory statement.” Could you be more specific? What kind of wordings should elicit what kind of responses from LVA? The samples we intend to use as “relevant segments” for evaluating LVA are linguistically neutral sentences which do not contain language that reveals the topic of the passage; nor do they contain affective words or phrases. An example would be “I have thought about this for some time and have come to a pretty firm conclusion.” These sentences are embedded within a passage that expresses views that contradict the subject’s strongly held beliefs. We have observed, from our independent physiological and behavioral measures, that the subjects we employed maintained their high levels of emotion while speaking these embedded phrases.

3. “Length of statement being read.” As you recall from the demonstration we provided you, subjects read a 5-8 sentence paragraph. We do not know exactly how to determine whether someone’s “interest” varied over the course of reading, although we concurrently measured galvanic skin response and pulse rate, which might be expected to be at the lower end of the speaker’s range of values if they were bored. However, for the research, we only used deceptive and stressed samples which showed a 50% or greater shift (usually much greater) from the speaker’s baseline (as measured by the physiological correlates of stress combined with two self-report scales). We can thereby demonstrate that our speakers were under a substantial degree of stress and therefore were engaged with the task when producing what we classify as deceptive samples and stressed samples.

4. “Consistent introduction for calibration purposes.” We are aware of this requirement from the course completed by two members of our research team. Thus, we intend to use the same standard passage, the “Rainbow Passage,” as read by each speaker for the calibration of each speaker’s individual samples and this includes all those submitted for LVA. Why would this procedure not robustly comply with the requirements that were included in our training?

5. “The reader’s particular relationship” and “Unique issue-related personal experience.” We deem these to be irrelevant issues also because it (again) is impossible for anyone to actually quantify specific nuances of the beliefs in question. For each deceptive and truthful passage read by a subject, we inquired as to whether or not all aspects of the passage fully constituted a lie/truth given their beliefs. Subjects who felt that even small portions of the passage were not a lie/truth were given a different one. If no passages were identified as deceptive or truthful, the subject was excluded from the project. As for “unique issue-related personal experience,” it is not possible to document every relevant memory in a person’s history that might be triggered in producing these particular lies and truthful statements. Therefore, this issue was not found to be even marginally relevant. As a matter of fact, it can be argued that, if this condition is an important one there is little chance that LVA can operate validly at all in the “real world.”

6. “We do not generally favor the use of reading in LVA examinations.” All of the samples used in the basic experiment involve read speech. Could you clarify your predictions for us? Will LVA show no sensitivity at all to deception in read speech? Are you saying that “cognitive effort” by the speaker (induced by reading in our case) interferes with the speech patterns that reveal that a speaker is lying? Would this not constitute an effective countermeasure for LVA’s deception detection capacity? In any event, we request a clarification in terms of LVA’s capacity to detect stress: will LVA show no sensitivity to stress in read speech? Indeed, is it possible that it simply would not be able to detect any of the relationships that you list under these conditions?

7. “Ratio of true to false statements.” As stated above, significant calibration material in the form of the Rainbow Passage is provided along with the short linguistically neutral sentence. The ratio in duration between the neutral sentences that can contain stress or deception) and the calibration passage is on the order of 1:3 to 1:5.

8. Recommendation 1, p. 3: We are planning to analyze only one deceptive sentence per speaker, so our protocol and your suggestion match.

9. Recommendation 2, p. 3: Please clarify what is meant by “lengthy.” Our individual wave files (one per speaker, containing the calibration passages and the relevant material) are typically 30-35 seconds in length. These durations would be longer than most encountered in “real life.”

10. Recommendation 3, p. 3: We do not deem this issue relevant or even applicable; the utterance that contains deception consists of a short, neutral sentence that does not contain topical information. The neutral sentence does not contain just one or two -- but

rather 17-23 words -- as we knew 1-2 word samples would be unanalyzable by LVA (from the training course you provided us). As for "context," we assume context refers to the information expressed over an entire recording session; if so, it could have serious implications for the interpretation of the information in the relevant portions. However, this issue is not applicable in our case since, by our intention, the relevant information appears out of context. It would be poor science indeed if we provided an LVA operator with contextual information about the relevant materials (e.g., those containing lies or stress). Rather, we judge that it is proper to test the performance of the product independent of its use by an operator. We are sure that you understand that we must do so if we are to see how accurate your product is in detecting deception (in the signal) independent of how good the operator is at listening to speech content and making judgments about whether or not the speaker was lying. To do otherwise would be to bypass your system.

11. "Due to reading . . . there may be a high JQ and stress throughout" p.3: Do you mean that all passages read by our subjects, including those in the low stress conditions, will show a high JQ (and please define high JQ -- your manual sets it at 35 and above to be above "normal"). If not, please state your predictions for JQ for our low stress conditions (i.e. reading low stress truthful statements and low stress lies) rather than high stress truthful statements (involving shock) and high stress lies (uttered with elevated stress levels due both to the nature of the lie and/or anticipation of shock).

12. "This may at times create random deception pop-up messages related to complicated words" p.4: Could you please be more specific? How often do you predict false positive deception responses from LVA and how often do they occur under the conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie)? Can you tell us how we can determine what constitutes a "complicated word" for a given subject?

13. "SPJ may be inconsistent from subject to subject . . . it might increase or decrease dramatically" p.4: Could you please be more specific and, in doing so, please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie)? What are your quantitative predictions?

14. "LJ and AVJ may be higher than the normal . . . but if the subject loses concentration . . . drop down significantly" p.4: Again these comments appear irrelevant at best, primarily because it is impossible for anyone to actually quantify, on a moment-to-moment basis, when a subject goes into an "auto-pilot" mode. As in similar statements you have made, any reference to a specific cognitive state or specific past experiences can not be verified by any known means; therefore, they cannot be used to evaluate your voice analysis product. If these variables must be measured and controlled for in order to test LVA, then it appears impossible to do so -- or, indeed, employ LVA for any meaningful purpose. Please advise if we are not interpreting your statements correctly.

15. "Fmain . . . may be unpredictable" p.4: Given this, we assume that these parameters should be ignored in the final protocol. We are amenable to that, as reflected in our

proposed protocol in which we examine JQ for stress and use the DI/NDI judgment to determine for deception.

16. Recommendation 1, p.4: Do you mean 25-35 for the JQ range? If so, please reconcile this with the JQ scale in your manual, which suggests 35 and above for abnormally high stress. In addition, could you please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) in making quantitative predictions for JQ?

17. Recommendation 2, p.4, "and it may not be present . . . emotional significance." This recommendation was not deemed applicable at all because it once again refers to unverifiable cognitive and/or emotional states. We cannot know whether or not particular subjects perceived shock to be insignificant. We do know that our subjects' stress levels increased dramatically in anticipation of shock. Please note also, that subjects who did not respond to shock (e.g., increase in GSR, pulse rate, self-report rating scales) were not included in the final database used to test LVA.

18. Recommendation 3, p.4: Could you please refer to conditions of our basic study protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) in making quantitative predictions for global stress response?

19. Recommendation 4, p.5: Again, could you please specifically refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) in making quantitative predictions for SPT by gender?

20. Recommendation 5, p.5: This recommendation was not judged applicable because any difference in reading ability within our literate population of subjects was not measured. While we did not employ subjects who were unable to read the passages in a reasonably fluent manner, we did not document any subtle differences in their (reading) ability. However, are you implying that LVA could not be used with people of low intelligence? In any case, our recommendation would be to exclude SPJ from analysis.

21. Recommendation 6, p.5: Could you please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) when making quantitative predictions for AVJ?

22. Recommendation 7, p.5: Could you please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) when making quantitative predictions for imagination?

23. Recommendation 9, p.5, "but many or most readers may not intend or conceive of the contrary statements as 'lying'." The recommendation does no appear applicable or even useful, because it refers to an absolutely unverifiable cognitive and/or emotional states. How can any examiner or instrument know when a subject is truly lying about their convictions versus simply making contrary statements without intending to lie? How can we "read their minds" if there are no methods we can use to do so? Our subjects were

instructed to lie about a belief they held dearly; they were instructed that they would be heard by their peers and people in their community; they were instructed to sound convincing as they lied; they claimed to comply with our instructions and their behaviors validated this relationship. If verification was not present, they were eliminated from the study.

24. Recommendation 9, p.5: Could you please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) when making quantitative predictions for high SOS? Also, please quantify "occasionally" in this prediction or it must be excluded from the protocol.

25. Recommendation 10, p.5: Could you please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) when making quantitative predictions for sub EMO?

26. Recommendation 11, p.5, "how much extra cognition will go into the reading process." This recommendation does not appear to be reasonable because it refers to unverifiable cognitive and/or emotional states.

27. Recommendation 12, p. 5, "rise in stress . . . SPT/Emotional." Could you please refer to conditions of our protocol (e.g., low stress truth, low stress lie, high stress truth, high stress lie) when making quantitative predictions for these parameters?

28. Recommendation 12, p.5, "a sharp decrease . . . anger response to the pain." This recommendation was not deemed applicable because it refers to an unverifiable cognitive and/or emotional state.

29. Recommendation 13, p.5: Please restate this recommendation in a more specific manner or exclude it from the protocol recommendations.

We look forward to your responses to our queries. The final (agreed on) version of the protocol must be a straightforward one. At the very least (i.e., for our deception/truth and stress/unstressed samples), we will examine and score the presence or absence of deception or stress as based on either a categorical judgment from LVA (e.g., DI, NDI) or on a threshold for one or more parameters (e.g., a JQ score of 35 or above). In essence, once we agree on a protocol, our job will be to extract the analysis provided by LVA. In essence, no human judgments will be involved. As researchers we will simply collate the results of LVA's analysis (as based on the instructions we received in your training program). If your team would like to analyze our database using a more "free-form" approach in which you use your judgment as operators to weigh a variable number of parameters to classify a sample as "deceptive" or "nondeceptive" (or "stressed" or "nonstressed"), we would we would be happy to provide the database to you -- as well as an answer sheet for scoring your results (your analysis would take place at our facility at a time convenient for you). If you provide us with your scores, we are willing to report on LVA's sensitivity to deception and to stress as used by your group.

One final comment. Whereas we can evaluate the LVA equipment on a straightforward lie/truth, stress/nonstress basis -- and, while we can do so using the personnel you have trained for us -- we are concerned about the vulnerability of your system for use in the real world. If all the points you make in your letter are true, we are a little apprehensive of trying to use the “real-life” materials that we promised you we would. We now wonder if even our rigorous approach in that regard would reveal any meaningful relationship re: the field materials.

We look forward to continue working with you.

Best regards,

J.D. Harnsberger, PhD
Assistant Professor

Harry Hollien, PhD
Professor Emeritus